Objective Bayesian Variable and Function Selection with Hyper-g Priors

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Daniel Sabanés Bové

aus

Deutschland

Promotionskomitee

Prof. Dr. Leonhard Held (Vorsitz)
Prof. Dr. Reinhard Furrer
Prof. Dr. Hans-Rudolf Künsch (ETH Zürich)

Zürich, 2013

Preface

Fortunately many people have helped and supported me during my PhD time. Now I would like to use the opportunity to thank them for their company and support.

In the first place I want to thank my supervisor Leonhard Held for his steady and strong encouragement during my statistics career, and the opportunity to write my PhD thesis under his excellent guidance and leadership. I was really lucky to meet him during his Munich time, and learning the Bayesian 101 from him. Following him to Zurich and being part of his world-class Biostatistics group was a most inspiring and pleasant experience. Thank you for always having an open door at your office so that I could come by and discuss the latest developments in research!

My fellow PhD colleagues were invaluable during the work on my thesis. Especially my office mates, Sebastian Meyer and Rafael Sauter (and previously Birgit Schrödle), contributed to the fun part of the work in the Institute for Social and Preventive Medicine. I wish I had more time to go swimming with you! Shared conference experiences with Julia Braun, Andrea Riebler (on both sides of the conference!), and Sarah Haile will always stay as good memories in my mind. I would also like to thank Julia for technical support with the thesis template and printing. Wei Wei supported us with excellent Chinese food and cake a lot of times, and contributed to the exciting international experience of my PhD time, together with Michaela Paul, Andrea Kraus, Lorenzo Tanadini and Steffi Muff. Thank you for the many good discussions, and I will miss the cake club, and going to the Mensa with you!

Sinikka Kohler helped me with all administration issues, including the start here in Zurich. Malgorzata Roos was always in good temper and very grateful for even small amounts of help I was happy to give, and I really enjoyed teaching the exercises of the block course on Bayesian Biostatistics with her. Eva Furrer also helped a lot with the block course, and organized nice Master study meetings. Burkhardt Seifert kept us up-to-date with consulting issues and the latest weather announcements. Alois Tschopp provided excellent and immediate support for any IT problems I had, including help with setting up our division server "Andan" (also a heritage of Andrea Riebler!) and last but not least printing this thesis.

I express my thanks to Gonzalo García-Donato who kindly accepted to review my thesis. Furthermore, I am grateful to Reinhard Furrer and Hans-Rudolf Künsch for being part of my dissertation committee.

Finally, I want to thank my wonderful wife Katja for all her love and constant support, and also for her incredible patience during many weekends and evenings of work on this thesis.

Zurich, September 2013

Daniel Sabanés Bové

Zusammenfassung

Die zwei grössten Herausforderungen der Bayesianischen Modellwahl sind die Spezifizierung von Priori-Verteilungen für die Parameter aller Modelle und die Berechnung der daraus resultierenden Posteriori-Wahrscheinlichkeiten der Modelle über die marginalen Likelihood-Werte. Mittlerweile gibt es eine breite Literatur zu automatischen und objektiven Priori-Verteilungen. Diese befreien den Statistiker von der manuellen Spezifizierung der Priori-Verteilungen für die Parameter, die schwierig ist wenn keine substantielle Priori-Information vorliegt. Ein wichtiger Vertreter ist die g-Priori von Zellner, die im linearen Modell aufgrund verschiedener günstiger Eigenschaften beliebt ist. Daraus entstehen stetige Mischungen von g-Priori-Verteilungen wenn man wiederum eine Priori-Verteilung für den Priori-Kovarianzmatrix-Faktor g annimmt. Diese sogenannten Hyper-g Priori-Verteilungen erübrigen die manuelle Wahl von g, das sehr einflussreich in der statistischen Analyse sein kann, und erhalten teilweise trotzdem eine geschlossene Form für die marginalen Likelihood-Werte.

In einer früheren Arbeit benutzten wir fraktionelle Polynome (FP), die eine Erweiterung der klassischen Polynome sind, in Verbindung mit Hyper-g Priori-Verteilungen, um Kovariablen- und Funktions-Wahl in linearen Modellen zu betreiben. Für generalisierte lineare Modelle (GLM) ist eine Normalverteilung mit Null als Mittelwertsvektor und mit g multiplizierter inverser erwarteter Fisher-Informations-Matrix als Kovarianzmatrix der natürliche Kandidat für eine verallgemeinerte g-Priori. Die verallgemeinerte Hyper-g Priori-Verteilung beinhaltet zusätzlich eine Priori-Verteilung für g. Wir lösen das Hauptproblem, die Berechnung der marginalen Likelihood-Werte, mittels einer integrierten Laplace-Approximation. Diese erlaubt eine effiziente Erkundung des Modellraums mittels einer stochastischen Modell-Suche basierend auf Markov-Ketten Monte Carlo, da sie die gleichzeitige Ziehung von unterschiedlich dimensionierten Parametern der verschiedenen Modelle vermeidet. Nachdem vielversprechende Modelle gefunden wurden, können jeweils die Parameter mit Hilfe eines Metropolis-Hastings Verfahrens gezogen werden.

Splines sind flexibler als FP und damit eine attraktive Alternative. Wir stellen sie als gemischte Modelle dar, wobei der nicht-lineare Anteil durch die zufälligen Effekte parametrisiert wird. Nachdem diese heraus integriert sind, können wir die Hyper-g Priori-Verteilung auf die verbliebenen Koeffizienten, welche die linearen Anteile der Kovariablen-Effekte parametrisieren, anwenden. Ein additives Modell ist dann definiert durch die (ganzzahligen) Freiheitsgrade aller Kovariablen-Effekte, wobei wir auch den Ausschluss von Kovariablen und exakt lineare Effekte zulassen. Für GLM verwenden wir den iterierten gewichteten Kleinste-Quadrate Algorithmus um ein lineares Modell zu erhalten, von dem wir dann die passende Struktur der Priori-Kovarianzmatrix für die Hyper-g Priori-Verteilung ableiten. Eine Simulationsstudie zeigt auf dass unser Verfahren konkurrenzfähig ist im Vergleich zu anderen Bayesianischen additiven Modellwahl-Verfahren. Wir verwenden es zur Schätzung des Diabetes-Risikos mittels logistischer Regression.

Um Überlebenszeiten zu analysieren, erweitern wir die Hyper-g Priori-Verteilung auf Proportionale Hazards Regression. Als ersten Ansatz verwenden wir eine Poisson-Approximation der vollen Likelihood, die bereits von Cai und Betensky (2003) vorgeschlagen wurde. Wir beschreiben wie diese fehlerhafte Approximation mit Hilfe einer Erweiterung des Datensatzes korrigiert werden kann. Diese Methode hat den Nachteil dass der Datensatz quadratisch mit der Stichprobengrösse wächst. Der zweite Ansatz erhält die lineare Daten-Komplexität und basiert auf sogenannten Test-basierten Bayes Faktoren (TBF), die von Johnson (2005) vorgeschlagen wurden. Statt die marginalen Likelihood-Werte für die Original-Daten zu berechnen, werden sie hier für die (partiellen) Likelihood-Quotienten Teststatistiken (auch als Devianzen bezeichnet) berechnet. Wir erklären wieso die implizit angenommene Priori-Verteilung

genau unserer verallgemeinerten g-Priori-Verteilung entspricht. Wir spezifizieren eine Priori-Verteilung für den Skalierungsfaktor g, was uns zu TBF-basierten Hyper-g Priori-Verteilungen führt. Bei der Entwicklung eines klinischen Vorhersage-Modells mit logistischer Regression beobachten wir eine gute Approximations- und Vorhersage-Genauigkeit unseres Ansatzes. Bei der Anwendung auf Cox-Regression erhalten wir ähnliche Ergebnisse wie mit der Poisson-Approximation.

Abstract

Bayesian model selection poses two main challenges: the specification of parameter priors for all models, and the computation of the resulting posterior model probabilities via the marginal likelihoods. There is now a large literature on automatic and objective parameter priors, which unburden the statistician from eliciting manually the parameter priors for all models in the absence of substantive prior information. One important example is Zellner's g-prior, which has become a favourite choice of prior in the Gaussian linear model, due to various favourable properties. Continuous mixtures of Zellner's g-priors are obtained by assigning a hyperprior to the prior covariance factor g. These hyper-g priors avoid the user's choice of g, which can be very influential in the statistical analysis, and allow for a closed form marginal likelihood for specific hyperpriors.

In earlier work we used fractional polynomial (FP) transformations, which are an extension of classical polynomials, together with hyper-g priors, to perform variable and function selection in Gaussian models. For generalized linear models (GLMs), a natural candidate for a generalized g-prior is a mean-zero Gaussian prior on the regression coefficients, with the inverse expected Fisher information multiplied with g as the covariance matrix. The generalized hyper-g prior specifies an additional (arbitrary) hyperprior on the scaling factor g. We solve the main difficulty, the computation of the marginal likelihood, with an integrated Laplace approximation. This accurate approach allows to explore the model space with a Markov chain Monte Carlo (MCMC) based stochastic search, avoiding the simultaneous sampling of model parameters of varying dimensions and yielding a sample of promising models. Subsequently we sample model-specific parameters using a tuning-free Metropolis-Hastings algorithm.

Splines are an attractive alternative to FPs, because they are more flexible. We represent the splines as mixed models, where the non-linear parts are parametrized by the random effects. After integrating them out, we can apply the hyper-g prior to the remaining coefficients that parametrize the linear parts of the covariate effects. Each additive model is defined by the collection of (integer) degrees of freedom for all covariates, where we also allow for exclusion and strictly linear inclusion of covariates. For GLMs, we use the the iteratively weighted least squares algorithm to obtain a linear model approximation, from which we then derive the appropriate form of the prior covariance matrix for the hyper-g prior. In a simulation study we find that our method performs competitively in comparison with several other Bayesian additive model selection procedures. We use the method to derive logistic regression models for estimating diabetes risk.

In order to analyse survival data, we extend the hyper-g prior to proportional hazards regression. The first idea is to use a Poisson model approximation of the full likelihood, which was first proposed by Cai and Betensky (2003). We describe how it can be corrected, and obtain a data augmentation which has quadratic complexity in the sample size. The second idea retains linear complexity, and builds on so-called test-based Bayes factors (TBFs), which were proposed by Johnson (2005). Instead of computing the marginal likelihood for the original data, it essentially computes the marginal likelihood for the (partial) likelihood ratio test

statistics (also called deviances). We explain that the prior which is implicit in this approximation is exactly our generalised g-prior, and assign a hyperprior to the scaling factor g, which leads to TBF-based hyper-g priors. For the development of a clinical prediction model with logistic regression, we observe good approximation accuracy and competitive performance in a bootstrap study. For a Cox regression application, we observe similar results as with the Poisson model approximation.

Thesis outline

Introduction

Paper I: **Hyper-g priors for generalized linear models**

Daniel Sabanés Bové & Leonhard Held

Paper published in *Bayesian Analysis*, 2011, **6**, 387–410.

Paper II: **Objective Bayesian model selection in generalised additive models with penalised splines**

Daniel Sabanés Bové, Leonhard Held & Göran Kauermann

Paper conditionally accepted and revised for *Journal of Computational and Graphical Statistics*.

Paper III: **Comment on Cai and Betensky (2003), On the Poisson approximation for hazard regression**

Daniel Sabanés Bové & Leonhard Held

Letter to the Editor published in *Biometrics*, 2013, **69**, 795.

Paper IV: **Approximate Bayesian model selection with the deviance statistic**

Daniel Sabanés Bové & Leonhard Held

Appendix I: **Hyper-g priors for generalised additive model selection**

Daniel Sabanés Bové, Leonhard Held & Göran Kauermann

Extended abstract published in the Proceedings of the 26^{th} *International Workshop on Statistical Modelling*, Valencia, Spain, 2011.

Appendix II: **Software manual**

Introduction

Almost all statistical inference is based on statistical models. Statistical models describe, in a rather abstract and mathematical way, how structure and randomness produce observable events. The models are defined by their structure and have model parameters that endow them with flexibility. Usually, a specific model is chosen by the researcher, based on conventions or on subject-matter considerations. After obtaining a set of observations, the crucial assumption is made that they were really generated in a way that can be described by the statistical model. The corresponding model parameters can then be optimised to adapt the model to the data set. This is the estimation of model parameters from data. The model is thereby fitted to the data set, and can now serve for different purposes. Hypotheses about the unknown true parameter values can be assessed in light of the parameter estimates and the uncertainty about them. The parameter estimates can be interpreted in the model framework for the reality. Last but not least, the fitted model can be used to predict future observations, with or without the availability of partial information about them.

Although model-based statistics is very successful, it depends heavily on the choice of the model. Cox (1990) notes that the "choice of an appropriate family may be the most challenging phase of analysis". This challenge is the topic of this thesis. Cox (1990) further distinguishes three different roles for statistical models: Substantive models are directly motivated by subject-matter considerations, and often specific to one application. Empirical models seek to capture associations in the data which may not be directly due to the application mechanics, this makes them more generally applicable. Indirect models are rather used as black boxes to summarize the data. This thesis is concerned with empirical models, more specifically general-purpose regression models, *i. e.* models that describe the conditional distributions of outcome observations ("response") based on known features ("covariates"). Typical important questions when specifying the model are which covariates to include in the model ("variable selection"), and in which functional way the covariates are included in the model ("function selection").

Section 1 introduces the considered model classes. The statistical tools that we develop in this thesis are part of the objective Bayesian model selection family, to which Section 2 is dedicated. The focus is on two major challenges in Bayesian model selection: First, specifying an objective prior distribution for the model parameters is important, and several approaches proposed in the literature are surveyed. Second, for implementing the approaches for model comparison in practice, often the huge number of possible models is an obstacle. In Section 3 we review a number of modern stochastic model search algorithms, which are tackling this problem.

1 Regression models

As Aldrich (2005, p. 401) writes, "[i]n the 1920s R. A. Fisher (1890–1962) created modern regression analysis out of two nineteenth century theories: the theory of errors of Gauss and the theory of correlation of Pearson." Fisher was leading "the last phase in the historical development of the Gauss linear model". What is standard statistical thinking today, was two innovations at his time: "the normal linear regression specification that, conditional on the x's, y is normally distributed with its expectation linear in the x's, and the notion that for inference the x values could be treated as fixed." We briefly set here the notation for the regression models considered in this thesis.

Linear regression The classical linear regression model assumes that the response variables y_i ($i = 1, \ldots, n$) are independent and normally distributed conditional on the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, with expectations $\eta_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}$ and identical variance σ^2. We can write this assumption as

$$\begin{aligned} y_i &\overset{ind}{\sim} \mathrm{N}(\eta_i, \sigma^2), \\ \eta_i &= \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}. \end{aligned}$$

Here β_0 is the intercept term, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the regression coefficients vector.

Generalized linear regression In the generalised linear regression model (GLM, see McCullagh and Nelder, 1989), the normal distribution of the response variables y_i is replaced by a member of the exponential family. This includes many important distributions such as the Poisson, binomial, negative-binomial and exponential distributions. GLMs can thus be applied to data with binary and count response as well as to data with strictly positive continuous response. The response function (or inverse link function) h transforms the linear predictor η_i to the mean $\mathbb{E}(y_i) = \mu_i = h(\eta_i)$, which in turn is mapped to the canonical parameter $\theta_i = (db/d\theta)^{-1}(\mu_i)$ of the distribution. Often the canonical response function $h = db/d\theta$ is used where $\theta_i = \eta_i$. Here $db/d\theta$ is the first derivative of the function b as defined in the likelihood contribution

$$\mathrm{p}(y_i \mid \beta_0, \boldsymbol{\beta}) \propto \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi_i} \right\}$$

of the ith observation. The dispersions $\phi_i = \phi / w_i$ can incorporate weights w_i. The variance $\mathrm{Var}(y_i) = \phi_i d^2 b / d\theta^2(\theta_i)$ is expressed through the variance function $v(\mu_i) = d^2 b / d\theta^2((db/d\theta)^{-1}(\mu_i))$ as $\mathrm{Var}(y_i) = \phi_i v(\mu_i)$.

Proportional hazards regression For survival data, the response is the survival time t_i. Cox (1972) introduced the most commonly used approach known today under the names Cox regression or proportional hazards regression. A hazard function $\lambda(t)$ can be defined through the density function $\mathrm{p}(t)$ and the survival function $\mathrm{S}(t) = 1 - \mathrm{F}(t) = 1 - \int_0^t \mathrm{p}(u)\, du$ as $\lambda(t) = \mathrm{p}(t) / \mathrm{S}(t)$. Here it is assumed that the hazard function for the ith individual is given by

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

which of course leads to hazards that are proportional between individuals, since $\lambda_i(t) / \lambda_j(t) = \exp\{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{\beta}\}$ is constant with respect to the time t.

The survival times t_i are often right-censored, which means that it is only known that death happened at a time larger than t_i. This has to be taken into account in the regression analysis. Typically censoring indicators δ_i are used, with $\delta_i = 1$ the survival time has been fully observed while for $\delta_i = 0$ the observation is censored. The log-likelihood function is then given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^{n} \Lambda_i(t_i), \tag{1}$$

where $\Lambda_i(t) = \int_0^t \lambda_i(u)\, du$ is the cumulative hazard for the ith individual. Since the baseline hazard $\lambda_0(t)$ is often not of interest in the application, the partial likelihood estimation approach (Cox, 1975) leaves $\lambda_0(t)$ unspecified and does not estimate it. Assuming that the

observations are ordered such that $t_1 \leq \cdots \leq t_n$, the partial log-likelihood function is given by

$$\sum_{i=1}^{n} \delta_i \left[\boldsymbol{x}_i^\top \boldsymbol{\beta} - \log \left\{ \sum_{j \in \mathcal{R}_i} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}) \right\} \right],$$

where $\mathcal{R}_i$ is the risk set at time t_i, *i.e.* the indices of the observations with (observed or censored) survival times larger than t_i.

2 Objective Bayesian model selection

What is an objective statistical analysis? As Berger and Berry (1988) write, "to acknowledge the subjectivity inherent in the interpretation of data is to recognize the central role of statistical analysis as a formal mechanism by which new evidence can be integrated with existing knowledge." This implies that even an objective Bayesian analysis is also subjective, because it relies on assumptions that cannot be verified until a certain extent. It is not even clear what an objective Bayesian analysis is, as Berger (2006) lists four different philosophical viewpoints on this terminology. In this thesis we mostly follow the third position, which is: "Objective Bayesian analysis is a convention we should adopt in scenarios in which a subjective analysis is not tenable." Whether it is the *best* method for the analysis (this is the second position) is mostly beyond our scope.

2.1 Variable and function selection

Fortunately, there is a consensus on what Bayesian model selection is, which we outline here in the context of the regression models from Section 1. This thesis focuses on the two most common model selection problems in regression, variable and function selection.

Variable selection refers to the choice of the covariates for the vectors $\boldsymbol{x}_i$. An initial set of p covariates with values $x_{i1}, \ldots, x_{ip}$ for the ith observation is given. Then we need to decide for each covariate with index k, whether it is included ($\gamma_k = 1$) in the or excluded from the model $\mathcal{M}$. This model $\mathcal{M}$ can thus be represented by the binary vector

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p) \tag{2}$$

of the p binary inclusion indicators γ_k. The linear predictor

$$\eta_i = \beta_0 + \sum_{k=1}^{p} \gamma_k x_{ik} \beta_k$$

of the regression model retains the linearity in the covariate values x_{ik}.

Function selection refers to the replacement of the linear effect $x_{ik}\beta_k$ of the kth covariate in the linear predictor (2.1) by a non-linear function $f_i(x_{ik})$. Two possible function classes that are used in this thesis are splines and fractional polynomials (FPs). Splines are smoothly joined piecewise polynomials, where smoothness is defined in terms of the continuity of the derivatives at the knots (*e.g.* Durrleman and Simon, 1989). In Paper II we are going to use a specific class, the O'Sullivan penalised splines (Wand and Ormerod, 2008). FPs are global nonlinear functions, and generalise the classical polynomials by including also fractional and negative powers as well as the natural logarithm (Royston and Altman, 1994). In Papers I and IV we are extending the Bayesian approach for FPs in linear models (Sabanés Bové and Held,

2011) to GLMs. Note that Strasak, Umlauf, Pfeiffer, and Lang (2011) compare the two function classes with respect to their properties and their performance in simulation studies.

The model space is a finite set of regression models, defined through the included covariates and their functional form in the linear predictor. Given a prior distribution on all model parameters $\boldsymbol{\theta}_j$ (here the intercept β_0, the regression coefficients vector $\boldsymbol{\beta}_j$ and possibly a variance parameter σ^2), the marginal likelihood of a model $\mathcal{M}_j$ ($j \in \mathcal{J}$) can be computed:

$$\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j) = \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_j, \mathcal{M}_j)\, \mathrm{p}(\boldsymbol{\theta}_j \mid \mathcal{M}_j)\, d\boldsymbol{\theta}_j. \tag{3}$$

Note that the parameters also depend on the model index.

Using (3), the Bayes factor (Kass and Raftery, 1995) between model $\mathcal{M}_j$ and the null model $\mathcal{M}_0$ that only contains the intercept β_0 in the linear predictor, can be defined:

$$\mathrm{BF}_{j,0} = \frac{\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)}{\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_0)}. \tag{4}$$

Usually the prior distribution on $\boldsymbol{\theta}_j$ is assumed to factor as

$$\mathrm{p}(\boldsymbol{\theta}_j \mid \mathcal{M}_j) = \mathrm{p}(\boldsymbol{\beta}_j \mid \beta_0, \sigma^2, \mathcal{M}_j) \cdot \mathrm{p}(\beta_0, \sigma^2),$$

such that the prior on β_0 and σ^2 is the same for all models. Under certain conditions and based on Invariance and Predictive Matching arguments, Bayarri, Berger, Forte, and García-Donato (2012, section 3) justify this factorisation. The corresponding prior density $\mathrm{p}(\beta_0, \sigma^2)$ may even be improper, *i. e.* it need not integrate to 1. The technical reason is that any constant in this density cancels in the Bayes factor (4). The explanation is that β_0 and σ^2 are common parameters to all models, in which case improper priors are allowed, see again Bayarri et al. (2012, section 3) for a formal justification. For β_0, often the prior $\mathrm{p}(\beta_0) \propto 1$ is specified. In this thesis, the terms "improper flat prior", "flat prior" and "locally uniform prior" are synonyms for this prior. However, the conditional prior distribution on the regression coefficients vector $\boldsymbol{\beta}_j$ must be proper for all models, because this parameter changes between models. Therefore, the arbitrary constants in improper prior densities would *not* cancel in the Bayes factor (4).

In taking into account the prior model probabilities $\mathrm{p}(\mathcal{M}_j)$, we can finally compute the posterior model probabilities

$$\mathrm{p}(\mathcal{M}_j \mid \boldsymbol{y}) = \frac{\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)\, \mathrm{p}(\mathcal{M}_j)}{\sum_{k \in \mathcal{J}} \mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_k)\, \mathrm{p}(\mathcal{M}_k)} = \frac{\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)\mathrm{BF}_{j,0}}{\sum_{k \in \mathcal{J}} \mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_k)\mathrm{BF}_{k,0}}. \tag{5}$$

These can now be used to select the *maximum a posteriori* (MAP) model, which scores the highest posterior model probability. Alternatively a Bayesian model average (BMA) of the models can be built, with model weights given by (5).

2.2 Parameter priors

The literature on parameter priors for objective Bayesian model selection is huge. Therefore we focus here on the publications connected directly with g-priors, which are not already discussed in detail in the papers of this thesis.

g-priors Zellner (1986) proposed the g-prior for the regression coefficients in the linear regression model. It uses the crossproduct of the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ to build the covariance matrix of the Gaussian prior for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \,|\, \sigma^2 \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}). \tag{6}$$

Here we assume that the columns of $\boldsymbol{X}$ have been centered around zero, to ensure orthogonality of $\boldsymbol{\beta}$ to the intercept β_0. In this thesis, we understand orthogonality between parameters in the sense that the Fisher information matrix is block diagonal, *i. e.* it contains zero entries for the between-parameter off-diagonal elements (see *e. g.* Kass and Wasserman, 1995). In the linear regression model, the requirement $\boldsymbol{X}^\top \mathbf{1} = \mathbf{0}$ ensures that the expected Fisher information matrix is block diagonal. Note, however, that for the generalised linear regression model, this is only true in case of assuming the null model with $\boldsymbol{\beta} = \mathbf{0}$. The parameters β_0 and $\boldsymbol{\beta}$ are then called "null-orthogonal" (Kass and Wasserman, 1995). Moreover, there seems to be no clear justification of this orthogonalisation procedure in the literature. An alternative is to only do the centering for the prior covariance matrix in (6), leaving it unchanged in the model $\boldsymbol{y} \sim \mathrm{N}_n(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, see García-Donato and Martinez-Beneito (2012, section 2.1).

The intercept β_0 and the regression variance σ^2 are usually assigned improper Jeffreys priors, $\mathrm{p}(\beta_0, \sigma^2) \propto \sigma^{-2}$, such that the complete parameter prior is $\mathrm{p}(\beta_0, \boldsymbol{\beta}, \sigma^2) = \mathrm{p}(\beta_0, \sigma^2)\,\mathrm{p}(\boldsymbol{\beta} \,|\, \beta_0, \sigma^2) \propto \sigma^{-2}\,\mathrm{p}(\boldsymbol{\beta} \,|\, \sigma^2)$. Note that $\mathrm{p}(\boldsymbol{\beta} \,|\, \beta_0, \sigma^2) = \mathrm{p}(\boldsymbol{\beta} \,|\, \sigma^2)$ does not depend on β_0 but only on σ^2 is another implicit assumption of the g-prior.

Historically, the g-prior has been called "Reference Informative Prior" (RIP) by Zellner (1983), who motivates the construction with an imaginary sample obtained from the same linear regression model, but with scaled variance $g\sigma^2$. Zellner (1983) also proposes a more informative version with specified prior means for $\boldsymbol{\beta}$ and σ^2, which is extended by Agliari and Parisetti (1988) to specified prior variances for the elements of $\boldsymbol{\beta}$. Zellner (1983) already noted that the prior covariance factor g in (6) has a large influence on the resulting shrinkage of the posterior mean vector from the ordinary least squares estimate $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ towards the prior mean vector. Therefore this parameter is either fixed at recommended values (*e. g.* Fernández, Ley, and Steel, 2001; Ley and Steel, 2009), or it is assigned a hyperprior distribution. An early special case is the Zellner and Siow (1980) prior, which arises when g is assigned an inverse-gamma prior $\mathrm{IG}(1/2, n/2)$. Other hyperpriors are studied by Liang, Paulo, Molina, Clyde, and Berger (2008) and Ley and Steel (2012), and the resulting marginal priors on $\boldsymbol{\beta}$ are often called hyper-g priors. Implementations are available in the `R`-packages `BMA` (Raftery, Hoeting, Volinsky, Painter, and Yeung, 2013) and `BMS` (Feldkircher and Zeugner, 2009) available on `CRAN`.

Extensions Bayarri and García-Donato (2007) develop an extension of the Zellner and Siow (1980) prior for testing general hypotheses in general linear models. Their conventional prior is a special case of the divergence based (DB) prior proposed by Bayarri and García-Donato (2008), which generalises the Zellner and Siow (1980) prior to other situations than the linear regression model. The DB prior is based on ideas by Jeffreys (1961) and is derived from divergence measures between the competing models. Martinez-Beneito, García-Donato, and Salmerón (2011) use an approximation of the DB prior as the conditional prior for the slope change parameters in joinpoint regression of Poisson response.

Bayarri et al. (2012) make an effort to summarise and formalise criteria that should be fulfilled by parameter priors for objective Bayesian model selection. Previously, a list with Jeffreys's *desiderata* was given by Berger and Pericchi (2001). Besides the requirement that the conditional prior distribution of the regression coefficients must be proper to ensure that the Bayes

factors are well defined, Bayarri et al. (2012) list model selection consistency and information consistency as basic criteria: A prior is consistent with respect to model selection, if the posterior probability of the model that generated the data converges to 1 for increasing sample size n. Their more general definition of information consistency than that given by Liang et al. (2008) is that the Bayes factor of an alternative *versus* the null model must go to infinity, whenever the corresponding likelihood ratio statistic goes to infinity for increasingly large data sets. A rather new criterion is the intrinsic prior consistency, which is defined by the convergence of the regression coefficients prior to a proper prior that is independent of the data (*e. g.* the design matrix $\boldsymbol{X}$ for the case of the g-prior). This limit prior is an intrinsic prior (Berger and Pericchi, 1996). Note that Casella, Girón, Martínez, and Moreno (2009) examine the consistency properties of intrinsic priors, and an implementation is available in the `R`-package `varSelectIP` (Womack, Gopal, V., Novelo, L., Casella, and G., 2013). The predictive matching criterion requires the predictive distributions of two different models to be close with respect to a suitable distance, if the sample size is very small. Hence if the information is too sparse to discern between the models, the resulting predictions should be very similar. Finally, the parameter prior should give results that are invariant under changes of the units of the response or covariates, and invariant under group transformations. Bayarri et al. (2012) propose the "robust prior" for linear model selection, which generalises the hyper-g and hyper-g/n priors of Liang et al. (2008), and also gives closed form Bayes factors in terms of the hypergeometric function (Abramowitz and Stegun, 1964, section 15.3). Model selection with this prior is implemented in the `R`-package `BayesVarSel` (Garcia-Donato and Forte, 2014) on `CRAN`.

Connected to the question if priors have predictive matching between models is the question whether the prior specifications are compatible across models. Consonni and Veronese (2008) give answers to this question. They list four main strategies for deriving a compatible prior distribution in a submodel. Marginalization is just integrating out the regression coefficients that are not part of the submodel from the joint prior distribution. Usual conditioning fixes the parameters at the null hypothesis in the conditional prior distribution. Since this is not invariant to the formulation of the condition, Dawid and Lauritzen (2001) propose Kullback-Leibler projection and Jeffreys conditioning as solutions. Consonni and Veronese (2008) use these strategies to derive g-prior distributions that are compatible across models.

However, we note that the g-prior (6) is already compatible with usual conditioning. That means, if we split the $p = p_1 + p_2$ regression coefficients as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, and then the derive the conditional prior distribution for $\boldsymbol{\beta}_1$, given the fact that $\boldsymbol{\beta}_2 = \mathbf{0}$, we obtain exactly the same prior distribution as in the corresponding submodel, namely $\boldsymbol{\beta}_1 \mid \sigma^2 \sim \mathrm{N}_{p_1}(\mathbf{0}, g\sigma^2(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1})$. This can easily be shown by applying the rule for deriving the conditional normal distribution, and then comparing the resulting covariance matrix with the formula for inverting block matrices. This is another attractive property of the g-prior, which also translates to hyper-g priors that use a hyperprior on g.

High-dimensional problems One problem of the g-prior is that it does not work for high-dimensional linear models which have more covariates p than observations n, because then the crossproduct $\boldsymbol{X}^\top \boldsymbol{X}$ is singular. Gupta and Ibrahim (2007) propose to regularize the crossproduct matrix as in ridge regression (Hoerl and Kennard, 1970) by adding a small constant λ to its diagonal elements. They recommend to choose λ between 0.5 and 1, and report that the resulting bias of the posterior means is not severe. Baragatti and Pommeret (2012) performed additional simulation studies for tuning λ, and applied the g-prior to probit regression models. Celeux, Anbari, Marin, and Robert (2012) performed simulation studies to compare the

performance under a low informative setting when p is almost equal to n on simulated and real datasets. They conclude that the Bayesian methods, including hyper-g priors, produce more parsimonious variable selection than frequentist regularization methods, with equivalent prediction performance.

Krishna, Bondell, and Ghosh (2009) extend the g-prior with a power parameter λ on the empirical covariance of the predictors. They proceed by starting with the singular value decomposition (SVD) $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{\Gamma} \boldsymbol{S} \boldsymbol{\Gamma}^\top$ and use then $\boldsymbol{\Gamma} \boldsymbol{S}^\lambda \boldsymbol{\Gamma}^\top$ instead of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ for the prior covariance matrix. The original g-prior is obtained by $\lambda = -1$, and the identity matrix corresponding to ridge regression is obtained with $\lambda = 0$. The power parameter λ can control the degree to which the coefficients of correlated predictors are smoothed towards ($\lambda > 0$) or away ($\lambda < 0$) from one another.

Another generalisation using the SVD is proposed by Griffin and Brown (2010). It is based on the correlated normal-gamma distribution for $\boldsymbol{\beta}$ proposed in the same paper, which is parametrized by the shape parameters of the gamma mixture distributions and a correlation matrix. This prior leads to simultaneous shrinkage of marginal effects and differences, therefore clustering of regression coefficients in a group is possible. When the shape parameters are chosen as $\lambda s_j^{-k/2}$ where $\boldsymbol{S} = \text{diag}\{s_1, \ldots, s_r\}$ contains the singular values, and the correlation matrix is chosen as $\gamma \boldsymbol{\Gamma}^\top \text{diag}\{s_1^{-k/2-b}, \ldots, s_r^{-k/2-b}\}$, then the following holds: $k = 0, b = 0$: corresponds to a ridge prior on $\boldsymbol{\beta}$, while $k = 1, b > 0$ corresponds to a g-prior with $p > n - 1$ extension and extra sparsity shrinkage. A continuous blending of the two approaches is possible by assigning standard uniform priors to both k and b.

Another problem in high-dimensional settings is unveiled by Johnson and Rossell (2010) and Johnson and Rossell (2012). They show that commonly used parameter priors, among them the g-priors, lead to inconsistent model selection. The reason is that they are all "local" prior densities, *i. e.* the prior density function at the null hypothesis values is positive. In case of the g-prior, it is even the mode of the prior distribution. The assumptions under which the g-prior is inconsistent are: $p > \mathcal{O}(\sqrt{n})$ covariates, standard regularity conditions on the design matrices and regularity conditions on models with one extra covariate. As a solution, they propose "non-local" priors which have exactly prior density zero at the null hypothesis value. One example are the product moment densities (pMOM), which are obtained by multiplying a standard prior with $\prod_{i=1}^{p} \beta_i^{2r}$ where $r \geq 2$. The methodology is implemented in the `R`-package `mombf` (Rossell, Cook, Telesca, and Roebuck, 2013) available from `CRAN`.

Kundu and Dunson (2014) describe extensions of mixtures of g-priors to linear regression models where the residual density is unknown. As a nonparametric prior distribution for this residual density, they use a Dirichlet process mixture of normal distributions.

An alternative to a purely Bayesian procedure could be to rely on a frequentist procedure like the lasso (Tibshirani, 1996) for fast pre-selection of $p' < n$ covariates, and only afterwards to use a g-prior approach on the resulting subset of the model space. In principle, this would not be necessary, because we could only use the g-prior approach and constrain the model to be of full rank, while searching through the whole set of $p > n$ covariates. Comparing the lasso, the g-prior, and the combination of lasso and g-prior, in the $p > n$ would be a very interesting area of both theoretical and computational statistical research.

2.3 Model priors

The literature on model priors $\text{p}(\mathcal{M}_j)$ is relatively small compared to the literature on parameter priors. Most specifications are not at the center stage of the corresponding publications.

The most commonly used model prior (*e.g.* George and McCulloch, 1993; Raftery, Madigan, and Hoeting, 1997) for variable selection uses independent and identical Bernoulli priors $\mathrm{B}(\pi)$ for the inclusion indicators γ_{jk}:

$$\mathrm{p}(\mathcal{M}_j) = \prod_{k=1}^{p} \pi^{\gamma_{jk}} (1-\pi)^{1-\gamma_{jk}}.$$

If we denote the number of covariates included in model $\mathcal{M}_j$ as $d_j = \sum_{k=1}^{p} \gamma_{jk}$, we have $d_j \sim \mathrm{Bin}(p, \pi)$. If we fix the prior inclusion probability at $\pi = 1/2$, we obtain the uniform model prior with $\mathrm{p}(\mathcal{M}_j) = 2^{-p}$. Note that while this is a non-informative prior on the models, it is rather informative on their dimension, because d_j is then binomial with mean $\mathbb{E}(d_j) = p/2$, so small and large values of d_j are relatively unlikely *a priori*. Using a uniform prior on π, *i.e.* $\pi \sim \mathrm{U}(0,1)$ produces a uniform prior on the model dimension d_j (Geisser, 1984). This idea can be generalised to a beta prior for π with mean π_0 and equal distribution among all covariate choices for a specific number of covariates (Sala-I-Martin, Doppelhofer, and Miller, 2004). This generates a beta-binomial distribution on the dimension d_j (Ley and Steel, 2009), see also Brown, Vannucci, and Fearn (1998). Clyde and George (2004) mention a few more elaborations on the binomial prior theme.

Most important in the model prior distributions with a hyperprior on π is that they are multiplicity-corrected. This is most clear for the uniform hyperprior, which preserves the marginal inclusion probability of $1/2$ for all covariates. However, the inclusion indicators γ_{jk} are now dependent. Scott and Berger (2010) explain that the intuition behind the multiplicity correction with the fact that the posterior distribution of π can then concentrate near the true value when the number of potential covariates p increases with the true number d_j of influential covariates held fixed. Also empirical Bayes estimation (*e.g.* Carlin and Louis, 2000) of π protects against the multiplicity of testing, that is inherent in the variable selection problem.

An interesting idea is presented by Dellaportas, Forster, and Ntzoufras (2012) who argue that the parameter prior and the model prior must be jointly specified. Specifically, they propose to use model prior probabilities of the form $\mathrm{p}(\mathcal{M}_j) \propto \mathrm{p}_0(\mathcal{M}_j)(n/g)^{d_j}$, where $\mathrm{p}_0(\mathcal{M}_j)$ is the baseline model probability and d_j is the dimension of model $\mathcal{M}_j$.

3 MCMC and stochastic search for model space exploration

In this thesis we focus solely on the following approach for model space exploration: First, we either compute exactly or we approximate analytically the marginal likelihood of the models. This has the advantage that we do not need to take into account parameter spaces of different dimensions during the model space exploration. Second, we explore this model space via MCMC methods. After finding a promising set of models, we sample the model parameters in a third step, only for this set of models.

As a side note, we mention that we could have taken a fundamentally different computational approach by relying on reversible jump MCMC (RJMCMC) methods (Green, 1995) instead. In RJMCMC the MCMC is performed on the joint space of models and parameters. The advantage is that the marginal likelihood of the models need not be computed or approximated; instead the model sampling frequencies are directly used as estimates of the posterior model probabilities. The disadvantages are: 1) The construction of well-performing proposal distributions is complicated because of the varying parameter dimensions. 2) In order to ob-

tain reliable results, the sampler should have converged, which is difficult to assess. 3) The computational time and implementation complexity could be higher than with the approach pursued in this thesis. RJMCMC publications which are relevant for the topic of this thesis comprise Denison, Mallick, and Smith (1998), Biller (2000), Han and Carlin (2001), Dellaportas, Forster, and Ntzoufras (2002), Ntzoufras, Dellaportas, and Forster (2003), Nott and Leonte (2004), Fouskakis, Ntzoufras, and Draper (2009) and Forster, Gill, and Overstall (2012).

Graphical model selection The first approaches to exploring a discrete model space with deterministic and MCMC search were applications to graphical models, where the edges between the vertices define the model.

As a solution to the very large number of models in the classical Bayesian model average, Madigan and Raftery (1994) propose to exclude models which score much worse than the best model with respect to posterior model probability, or which score worse than a sub-model. The latter principle is based on Occam's razor (see *e. g.* Blumer, Ehrenfeucht, Haussler, and Warmuth, 1987), which lends the name "Occam's Window" for the averaging method. Madigan and Raftery (1994) design a deterministic search algorithm, based on ideas by Edwards and Havránek (1985).

Madigan and York (1995) then introduced the first stochastic search, the "Markov chain Monte Carlo model composition" (also abbreviated as MC^3), for graphical models. Given a model $\mathcal{M}$ and a suitably defined set of models $\mathcal{N}(\mathcal{M})$ in the neighbourhood of $\mathcal{M}$, the next model $\mathcal{M}' \in \mathcal{N}(\mathcal{M})$ is proposed. All models in the neighbourhood have the same probability to be selected, *i. e.* the proposal kernel is

$$q(\mathcal{M}' \mid \mathcal{M}) = \begin{cases} \frac{1}{|\mathcal{N}(\mathcal{M})|}, & \mathcal{M}' \in \mathcal{N}(\mathcal{M}) \\ 0, & \mathcal{M}' \notin \mathcal{N}(\mathcal{M}), \end{cases}$$

such that the acceptance probability in the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Hastings, 1970) equals

$$\min\left\{1, \quad \frac{\mathrm{p}(\mathcal{M}' \mid \boldsymbol{y})}{\mathrm{p}(\mathcal{M} \mid \boldsymbol{y})} \frac{|\mathcal{N}(\mathcal{M})|}{|\mathcal{N}(\mathcal{M}')|}\right\}.$$

Note that the ratio

$$\frac{\mathrm{p}(\mathcal{M}' \mid \boldsymbol{y})}{\mathrm{p}(\mathcal{M} \mid \boldsymbol{y})} = \frac{\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}')\,\mathrm{p}(\mathcal{M}')}{\mathrm{p}(\boldsymbol{y} \mid \mathcal{M})\,\mathrm{p}(\mathcal{M})}$$

is known, which makes MCMC feasible, even without knowing the normalising constant

$$\mathrm{p}(\boldsymbol{y}) = \sum_{j \in \mathcal{J}} \mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j) \tag{7}$$

of the posterior model distribution. York, Madigan, Heuch, and Lie (1995) apply MC^3 in a problem with missing data.

Madigan, Raftery, York, Bradshaw, and Almond (1994) compare Occam's Window with MC^3 in prediction examples using the logarithmic scoring rule, and find that both methods outperform any single model. Moreover, MC^3 performed better than Occam's Window.

Variable selection A quite popular version of MC^3 for variable selection problems is described (among others) by Brown et al. (1998) and Brown, Vannucci, and Fearn (2002) for

multivariate linear regression: From the current model $\mathcal{M}$, either a variable is added, or a variable is deleted, or a variable is exchanged. The last possibility is often called the "swap" proposal, because in the binary representation (2) of the model $\mathcal{M}$, a 0 and a 1 swap their places in the vector γ. The "swap" proposal is a way to avoid being trapped in one local mode of a multi-modal model posterior, which can easily occur when two covariates are highly correlated. Obviously there is considerable flexibility in customising the algorithm, *e.g.* via the specification of the proposal type probabilities. The algorithm is applied to multinomial probit models by Sha, Vannucci, Tadesse, Brown, Dragoni, Davies, Roberts, Contestabile, Salmon, Buckley, and Falciani (2004) using data augmentation with latent variables (Albert and Chib, 1993). Denison et al. (1998) apply a similar algorithm to knot selection for Bayesian spline curve fitting. They call the different proposal types the "birth", "death" and "move" steps. They combine the Metropolis-Hastings step with a Gibbs sampling step to draw the regression variance σ^2. See also Denison, Holmes, Mallick, and Smith (2002) for more examples.

Instead of a random walk proposal, Casella and Moreno (2006) use an independence proposal for their Metropolis-Hastings algorithm on the variable selection model space. Their proposal factors in the distribution of the number of included variables (*i.e.*, the model dimension), and the drawing of a model with the required number of variables. They initialise the model dimension distribution by sampling uniformly a fixed proportion of models of each possible dimension, and calculating the dimension probabilities that would result from truncating the model space to these models. Note that this "renormalisation" strategy is further discussed below. Given the dimension, the selection of covariates is drawn uniformly from all possible choices. When a new model is proposed, the dimension distribution is updated accordingly. Hence, the independence proposal adapts to the posterior model distribution during the course of the MCMC.

A quite complicated sampling scheme is proposed by Liang and Wong (2000). Their Evolutionary Monte Carlo (EMC) algorithm works by simulating a population of Markov chains in parallel, where a different temperature is attached to each chain, similarly to parallel tempering (Geyer, 1991). The population is updated by mutation, crossover and exchange operators, which are motivated by the genetic algorithm (Holland, 1975), a general optimisation technique mimicking the natural evolutionary process of chromosomes. The updates are accepted or rejected according to the Metropolis rule. They show in examples that their algorithm outperforms classical MCMC algorithms, both for sampling models as well as for finding the best models. Bottolo and Richardson (2010) extend EMC by including additional moves, sampling the g of hyper-g priors, and automatic tuning of the temperatures during the burn-in phase, and call the resulting algorithm Evolutionary Stochastic Search (ESS).

Theoretical analysis of the performance of the MCMC algorithms for model space exploration is difficult and therefore mostly neglected. However, it is a field of interest *e.g.* in computer science, see Jerrum and Sinclair (1996) as a starting point.

Sampling frequencies versus renormalised probabilities After running the Markov chain $\{\mathcal{M}^{(1)}, \mathcal{M}^{(2)}, \ldots, \mathcal{M}^{(T)}\}$ of length T through the model space and obtaining a unique set of models $\hat{\mathcal{J}}_T$, the question is how to proceed: should one use the sampling frequencies

$$\hat{\mathrm{p}}_{\mathrm{freq}}(\mathcal{M}_j \mid \boldsymbol{y}) = \sum_{t=1}^{T} \mathbb{I}(\mathcal{M}_j = \mathcal{M}^{(t)})/T \tag{8}$$

or rather the renormalised probabilities

$$\hat{\mathrm{p}}_{\mathrm{norm}}(\mathcal{M}_j \mid \boldsymbol{y}) = \frac{\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j)}{\sum_{j \in \hat{\mathcal{J}}_T} \mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j)} \tag{9}$$

to estimate integrals

$$\mathbb{E}(\Delta \mid \boldsymbol{y}) = \sum_{j \in \mathcal{J}} \Delta(\mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j \mid \boldsymbol{y}) \tag{10}$$

for quantities of interest Δ via replacing the true unknown posterior model probabilities $\mathrm{p}(\mathcal{M}_j \mid \boldsymbol{y})$? For example, Madigan and York (1995) use the sampling frequencies $\hat{\mathrm{p}}_{\mathrm{freq}}(\mathcal{M}_j \mid \boldsymbol{y})$. Recently, there is increased interest in comparing the two different approaches of processing the model sampling output.

Clyde and Ghosh (2012) "prove that renormalization of posterior probabilities over the set of sampled models generally leads to bias that may dominate mean squared error". They propose ratio Horvitz-Thompson estimators (Horvitz and Thompson, 1952) for (10), where the numerator $\sum_{j \in \mathcal{J}} \Delta(\mathcal{M}_j)\,\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j)$ and the denominator $\sum_{j \in \mathcal{J}} \mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j)$ of (10) are estimated separately. The resulting integral estimate is approximately unbiased, and simulation studies suggest that it yields a smaller mean squared error than the estimate obtained with the sampling frequencies (8) or the renormalised probabilities (9). Their method involves running a second independent Markov chain to estimate the normalising constant (7) (George and McCulloch, 1997), based on which then the probability of visiting the model $\mathcal{M}_j$ during T iterations, $1 - \{1 - \mathrm{p}(\mathcal{M}_j \mid \boldsymbol{y})\}^T$, is estimated as input for the Horvitz-Thompson estimators.

In a similar effort, García-Donato and Martinez-Beneito (2012) show that the sampling frequencies yield consistent estimators, while the renormalised probabilities yield biased estimators. In extensive empirical studies with the ozone data set introduced into the model selection literature by Breiman and Friedman (1985), they find that the sampling frequencies outperform the renormalised probabilities for estimating typical quantities (10), such as predictive expectations and variable inclusion probabilities. As objective parameter prior for the linear regression models, they use Zellner's g-prior (Zellner, 1986).

Strict search algorithms One advantage of the renormalised estimates (9) is that they do not rely on the fact that the model sampling method converges to the true posterior model distribution. This opens the door for algorithms that do not have a stationary distribution, and hence can search the model space more aggressively for models with high posterior probability.

In variable selection problems, Berger and Molina (2005) propose to guide the search direction by the variable inclusion probabilities. Each search iteration starts with sampling one of the models visited so far, with renormalised sampling probabilities. Afterwards, the algorithm decides to add or delete a covariate from the model with probability $1/2$, and samples from the corresponding set of covariates with probabilities being proportional to the odds of their inclusion or exclusion, respectively. Thereby, the algorithm only visits models differing in one parameter, which can be used for efficient updating of the posterior model probabilities. In line with this idea, Berger and Molina (2005) propose path-based priors, which are a variant of the Intrinsic Bayes factor (Berger and Pericchi, 1996).

With very similar ideas, Scott and Carvalho (2008) build their search algorithm called FINCS (feature-inclusion stochastic search), which is motivated by Gaussian graphical model selection. Here the inclusion probabilities for edges between vertices are updated on-the-fly and

used for navigating the search. In addition to the local moves described by Berger and Molina (2005), they include global moves, which shall avoid being trapped in one mode of a multimodal posterior model distribution. They compare their method with Gibbs sampling (George and McCulloch, 1993) and the algorithm proposed by Jones, Carvalho, Dobra, Hans, Carter, and West (2005) (see below), and find that their method finds better models than the two competing approaches. The FINCS algorithm is also applied by Carvalho and Scott (2009).

With renormalised estimates, each model does not need to be visited more than once. This fact is exploited by Clyde, Ghosh, and Littman (2011) for their Bayesian Adaptive Sampling (BAS) algorithm, which is designed for variable selection. After initialising the inclusion probabilities ρ_k ($k = 1, \ldots, p$), *e.g.* based on minimum Bayes factors (Sellke, Bayarri, and Berger, 2001) from the full model, the models are sampled without replacement from the distribution $\mathrm{p}(\mathcal{M}_j) = \prod_{k=1}^{p} \rho_k^{\gamma_{jk}} (1-\rho_k)^{(1-\gamma_{jk})}$, where each model $\mathcal{M}_j$ is defined by the p binary variable inclusion indicators γ_{jk}. The inclusion probabilities ρ_k are updated periodically during the sampling process based on renormalised estimates from the previously sampled models. Clyde et al. (2011) encode each model as a path in a binary tree, where in each level of the tree the decision whether to include the corresponding variable is encoded. After a model is sampled, only the conditional sampling probabilities along the model path need to be updated to truncate the distribution to the set of models not sampled so far. The BAS algorithm is implemented in the `R`-package `BAS` (Clyde, 2012).

Ma (2012) develops a related idea called Bayesian recursive variable selection, based on the representation of any prior or posterior model distribution as a forward-stepwise distribution. The corresponding prior sampling routine recursively adds covariates to the starting model, where the covariates are drawn with model-specific sampling probabilities, until a stopping random variable signals that the model is complete. The posterior model probabilities can be calculated by a sequence of recursions, which is exploited to adopt approximate recursive computation methods for trees to obtain an efficient search algorithm.

Heaton and Scott (2010) provide a review of Bayesian linear model selection approaches, and compare different stochastic search algorithms in examples, among them the ozone data set. They conclude that when searching for models with high posterior probability, true search algorithms are to be preferred over MCMC, while for the estimation of inclusion probabilities, they are pessimistic and do not give a clear recommendation.

Algorithms for parallel computing With the advent of parallel computing facilities in research and industry, both in supercomputers encompassing many nodes but also in laptops with multiple processors, the need to adequately exploit this computing power increases. Bayesian model search is no exception. Running multiple Markov chains in parallel is of course the most immediate but also a very naive approach to speed up the model search.

The Shotgun Stochastic Search (SSS) algorithm proposed by Hans, Dobra, and West (2007) is a more sophisticated approach. Starting from the current model $\mathcal{M}$, it starts by calculating in parallel the scores (unnormalised posterior model probabilities) of all models in the neighbourhood $\mathcal{N}(\mathcal{M}) = \mathcal{N}^-(\mathcal{M}) \cup \mathcal{N}^0(\mathcal{M}) \cup \mathcal{N}^+(\mathcal{M})$. Since many models are evaluated in the neighbourhood, this step gives the "shotgun shot" in the algorithm. Then one model each is sampled from the "deletion" moves $\mathcal{N}^-(\mathcal{M})$ which have one covariate less than the original model $\mathcal{M}$, the "replacement" moves $\mathcal{N}^0(\mathcal{M})$ and the "addition" moves $\mathcal{N}^+(\mathcal{M})$. Afterwards, one of the three resulting models is sampled. The sampling probabilities are based on the renormalised scores in each step. Of course, the application of SSS is not limited to variable selection in regression models. All that is required to apply the algorithm to another problem is the definition of neighbouring models. For example, Jones et al. (2005) use SSS to infer

Gaussian graphical models, where neighbouring graphs differ in one edge. Implementations are available at `http://isds.duke.edu/research/software`.

For ESS (Bottolo and Richardson, 2010), an efficient `C++` implementation is available in the software `GUESS` (Bottolo, Chadeau-Hyam, Hastie, Langley, Petretto, Tiret, Tregouet, and Richardson, 2011; Liquet, Chadeau-Hyam, Bottolo, Campanella, and Richardson, 2013). It offers the possibility to re-route computationally intensive linear algebra operations towards the Graphical Processing Unit (GPU), thus exploiting the graphics card parallel computing power of today's personal computers. However, GPU computing involves an overhead due to the data transfer between the Central Processing Unit (CPU) memory and the GPU memory. Therefore, this option is only recommended for large enough data sets. The authors also provide an `R`-package `R2GUESS` (Liquet and Chadeau-Hyam, 2014) which interfaces the `C++` library.

Schäfer and Chopin (2013) develop a sequential Monte Carlo algorithm for sampling from large binary spaces, and apply it to the special case of variable selection problems. The algorithm alternates importance sampling steps, resampling steps and Markov chain transitions, to recursively approximate a sequence of multivariate binary distributions, using a set of weighted "particles" which represent the current distribution. These distributions adapt then over the iterations to the true posterior distribution. One advantage of sequential Monte Carlo algorithms over traditional MCMC algorithms is that they can be parallelised in the sense that one can simulate the particles in parallel. Schäfer and Chopin (2013) report that they have processed variable selection problems from genetics with thousand covariates within a few hours, using a parallelised version of the algorithm on a cluster with 64 CPUs. A complete `Python` implementation is available at `http://code.google.com/p/smcdss`. See Durham and Geweke (2013) for GPU-based speed-up of sequential Monte Carlo.

Thesis Summary

This thesis consists of four papers. Their content and contribution are briefly summarized below. Appendix I presents an early version of Paper II, and Appendix II gives an introduction to the software implementing the approaches.

Paper I

Hyper-g priors for generalized linear models by Daniel Sabanés Bové and Leonhard Held.

This paper extends Zellner's g-prior (Zellner, 1986) from the linear model to generalized linear models (GLMs). Moreover, the hyper-g priors proposed by Liang et al. (2008) for the linear model are extended to GLMs by assigning a hyperprior to the hyperparameter g. Related approaches from the literature are summarized in a review section. An accurate calculation of the resulting marginal likelihood is achieved with an integrated Laplace approximation. A higher-order Laplace approximation can optionally be used. For sampling model-specific parameters, a tuning-free Metropolis-Hastings sampler is proposed. The approach is illustrated with variable and fractional polynomial (FP) selection in logistic regression modelling of the Pima Indian diabetes data.

This work is based on the idea by L. Held and me that our previous work on FP selection in linear models needed to be generalized to GLMs. I had the idea that the g-prior should again be a normal distribution, where just the prior covariance matrix has a different shape compared to the linear model case. I applied and implemented the INLA methodology from

Rue, Martino, and Chopin (2009) to calculate the resulting marginal likelihood, and used the implicit approximation of the posterior density for g as a proposal density in the Metropolis-Hastings sampler of the model parameters. After presenting the paper at the ISBA 2010 conference, I drafted the paper with illustrating applications, on which L. Held commented before I finalized it.

The main contribution of this paper is the extension of the hyper-g priors to GLMs, including a stable and accurate implementation in the R-package `glmBfp`, which is available on R-Forge. Besides, it contains a novel combination of the INLA methodology with Markov chain Monte Carlo (MCMC) to perform efficient posterior inference.

Paper II

Objective Bayesian model selection in generalised additive models with penalised splines by Daniel Sabanés Bové, Leonhard Held and Göran Kauermann.

This paper extends the hyper-g priors to generalised additive models. The additive covariate effects are modelled with penalised splines, represented as mixed models. After integrating out the random effects parametrising the non-linear parts of the splines, the hyper-g prior is applied to the fixed effects parametrising the linear parts. Each additive model is defined by the collection of (integer) degrees of freedom for all covariates, which derive from the random effects variances. A suitable objective model prior and a stochastic search algorithm are proposed. The methodology is first introduced for models for Gaussian response, and a simulation study demonstrates the advantages over other Bayesian model selection approaches. Consistently with the approach for GLMs, the methodology is extended to models for non-Gaussian response, and illustrated in a logistic regression application. The idea of meta-models differing only in the degrees of freedom of included covariates allows to define intuitive Bayesian model averages.

This work is based on the idea by G. Kauermann to specify the hyper-g prior for the fixed effects in the marginal model, and to discretise the model space by allowing only a finite set of degrees of freedom for the covariate effects. I started the manuscript from the initial draft by G. Kauermann, and generalised his idea to comprise any objective parameter prior for linear models. I implemented the method in the R-package `hypergsplines`, and wrote a separate R-package `appell` that calculates Appell's F1 hypergeometric function that is required for use of the hyper-g/n prior. L. Held had the idea to extend the objective model prior from variable selection. After G. Kauermann wrote the derivation of the approximate Fisher information for generalised additive models now contained in Appendix B of the paper, I discovered a simpler and more direct explanation based on the iteratively weighted least squares algorithm. I conducted the simulation study on the supercomputer "Schrödinger" and the logistic regression analysis. Both L. Held and G. Kauermann commented on the paper, which I subsequently finalized.

The main contribution of this paper is the idea to apply objective parameter priors developed for linear models to generalised additive models with penalised splines. The implementation with hyper-g priors shows advantages in a simulation study over competing approaches.

Appendix I presents an early version of Paper II which is published in the Proceedings of the 26^{th} International Workshop on Statistical Modelling (2011). This version is less detailed and does not include *e. g.* the simulation study and the meta-model idea.

Paper III

Comment on Cai and Betensky (2003), On the Poisson approximation for hazard regression by Daniel Sabanés Bové and Leonhard Held.

This Letter to the Editor contains a correction to the Poisson approximation for proportional hazards regression models proposed by Cai and Betensky (2003, section 5.1) and based on the univariate approach in Cai, Hyndman, and Wand (2002). It is shown that the original approximation, which uses a pseudo data set of the same size n as the original data set, and the resulting log-likelihood has an $\mathcal{O}(n)$ error. The correct approximation requires a pseudo data set that grows with n^2.

An extended version is attached. It describes in detail the employed trapezoidal cubature approximation to the cumulative baseline hazard, and slightly improves the original proposal by Cai et al. (2002). It contains an algorithm for computing the required offsets, which also accommodates data sets with ties between the survival times. An application example shows that the error might change the conclusions drawn from the statistical analysis.

The idea for this Letter to the Editor arose from the aim to extend the hyper-g priors from Gaussian, logistic and Poisson regression models covered so far to Cox regression models. After I obtained strange results with the original Poisson approximation, I discovered the error in its derivation and derived the correct approximation. I first drafted the extended version of the manuscript, on which L. Held commented. The Editor of Biometrics then asked for a shortened version suitable for a Letter to the Editor. After L. Held commented on my draft, I finalized it.

The main contribution of this Letter to the Editor is the exposure of the error in the initial publication of the Poisson approximation, and the correction of it, which can be used with a simple `R`-script written by me. Basically all Cox regression models can be fitted with this Poisson approximation, which also allows to estimate the baseline hazard.

Paper IV

Approximate Bayesian model selection with the deviance statistic by Daniel Sabanés Bové and Leonhard Held.

This paper merges the hyper-g prior methodology with the test-based Bayes factors (TBFs) proposed by Johnson (2005, 2008). Note that in this paper, the Bayes factor as defined in the Introduction is called data-based Bayes factor, in order to differentiate it from the TBF. It shows that if the deviance statistic is used for the TBFs, the generalized g-prior from Paper I is implicitly used. The TBF is a closed form expression in the hyperparameter g, the deviance and the dimension of the model, which allows to conveniently study the influence of g on shrinkage of regression coefficients and model selection. The paper reveals connections of empirical Bayes estimates of g to minimum Bayes factors and shrinkage estimates from the literature. As an alternative, fully Bayes estimation of g is proposed, which effectively implements TBF-based hyper-g priors. This approach is especially attractive for large model selection problems because of its computational efficiency, and the corresponding implementation issues are discussed in a separate section of the paper. As an example for the development of a clinical prediction model, variable and function selection in a logistic regression application is performed with the proposed test-based and the standard data-based Bayes factors. The competitiveness of the resulting predictions is evaluated with a bootstrap

study. The second data example illustrates the application to Cox regression, and shows that the results are close to those obtained with the Poisson approximation from Paper III.

The initial idea on which this paper is based was from L. Held, who suspected a relation between the g-priors and the TBFs, and proposed to specify a prior distribution for g. I found that the incomplete inverse-gamma prior (Cui and George, 2008) is conjugate to the TBF, and implemented the numerical integration required for non-conjugate hyperpriors. I proved that the generalized g-prior is implicitly assumed in the TBF derivation. I implemented the methodology in the `R`-package `glmBfp`, and conducted all analyses. While initially we hoped that we could extend the methodology to additive models with penalised splines, I found that this was not possible, and wrote the discussion section on this issue. I drafted the manuscript and L. Held commented on it and wrote the introduction. Afterwards we jointly finalized the manuscript.

The main contribution of this paper is the connection of the hyper-g priors with TBFs, resulting in a better understanding and estimation of the hyperparameter g and a more efficient approximation of the Bayes factors. Moreover it is the first usable approach to apply hyper-g priors to Cox regression.

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth Dover printing, tenth GPO printing edition, 1964.

A. Agliari and C. C. Parisetti. A-g Reference Informative Prior: A Note on Zellner's g Prior. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 37(3):271–275, 1988.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

J. Aldrich. Fisher and regression. *Statistical Science*, 20(4):401–417, 2005.

M. Baragatti and D. Pommeret. A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics and Data Analysis*, 56(6):1920–1934, 2012.

M. Bayarri and G. García-Donato. Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, 94(1):135–152, 2007.

M. J. Bayarri and G. García-Donato. Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(5):981–1003, 2008.

M. J. Bayarri, J. O. Berger, A. Forte, and G. García-Donato. Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics*, 40(3):1550–1577, 2012.

J. O. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

J. O. Berger and D. A. Berry. Statistical analysis and the illusion of objectivity. *American Scientist*, 76(2):159–165, 1988.

J. O. Berger and G. Molina. Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica*, 59(1):3–15, 2005.

J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

J. O. Berger and L. R. Pericchi. Objective Bayesian methods for model selection: introduction and comparison. In P. Lahiri, editor, *Model Selection*, volume 38 of *IMS Lecture Notes*, pages 135–207. Institute of Mathematical Statistics, Beachwood, OH, 2001.

C. Biller. Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics*, 9(1):122–140, 2000.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's Razor. *Information Processing Letters*, 24(6):377–380, 1987.

L. Bottolo and S. Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010.

L. Bottolo, M. Chadeau-Hyam, D. I. Hastie, S. R. Langley, E. Petretto, L. Tiret, D. Tregouet, and S. Richardson. ESS++: a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics*, 27(4):587–588, 2011.

L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

P. J. Brown, M. Vannucci, and T. Fearn. Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, 12(3):173–182, 1998.

P. J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):519–536, 2002.

T. Cai and R. A. Betensky. Hazard regression for interval-censored data with penalized spline. *Biometrics*, 59(3):570–579, 2003.

T. Cai, R. J. Hyndman, and M. P. Wand. Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics*, 11(4):784–798, 2002.

B. P. Carlin and T. A. Louis. Empirical Bayes: past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289, 2000.

C. M. Carvalho and J. G. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512, 2009.

G. Casella and E. Moreno. Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167, 2006.

G. Casella, F. J. Girón, M. L. Martínez, and E. Moreno. Consistency of Bayesian procedures for variable selection. *Annals of Statistics*, 37(3):1207–1228, 2009.

G. Celeux, M. E. Anbari, J.-M. Marin, and C. P. Robert. Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7 (2):477–502, 2012.

M. Clyde. *BAS: Bayesian adaptive sampling for Bayesian model averaging*, 2012. R package version 1.0.

M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

M. A. Clyde and J. Ghosh. Finite population estimators in stochastic search variable selection. *Biometrika*, 99(4):981–988, 2012.

M. A. Clyde, J. Ghosh, and M. L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101, 2011.

G. Consonni and P. Veronese. Compatibility of prior specifications across linear models. *Statistical Science*, 23(3):332–353, 2008.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, Aug. 1975.

D. R. Cox. Role of models in statistical analysis. *Statistical Science*, 5(2):169–174, 1990.

W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008.

A. P. Dawid and S. L. Lauritzen. Compatible prior distributions. In E. I. George, editor, *Bayesian Methods With Applications to Science, Policy and Official Statistics*, pages 109–118, Luxembourg, 2001. Eurostat.

P. Dellaportas, J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002.

P. Dellaportas, J. J. Forster, and I. Ntzoufras. Joint specification of model space and parameter space prior distributions. *Statistical Science*, 27(2):232–246, 2012.

D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(2):333–350, 1998.

D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Wiley, Chichester, 2002.

G. Durham and J. Geweke. Adaptive sequential posterior simulators for massively parallel computing environments. Technical report, University of Technology Sydney, 2013. URL `http://www.quantosanalytics.org/garland/gpu2.pdf`.

S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561, 1989.

D. Edwards and T. Havránek. A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72(2):339–351, 1985.

M. Feldkircher and S. Zeugner. Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in Bayesian model averaging. Technical Report 09/202, IMF, 2009.

C. Fernández, E. Ley, and M. F. J. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

J. Forster, R. Gill, and A. Overstall. Reversible jump methods for generalised linear models and generalised linear mixed models. *Statistics and Computing*, 22(1):107–120, 2012.

D. Fouskakis, I. Ntzoufras, and D. Draper. Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Annals of Applied Statistics*, 3(2):663–690, 2009.

G. Garcia-Donato and A. Forte. *BayesVarSel: Bayesian variable selection in linear models*, 2014. URL `http://CRAN.R-project.org/package=BayesVarSel`. R package version 1.4.

G. García-Donato and M. A. Martinez-Beneito. On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2012.

S. Geisser. On prior distributions for binary trials. *The American Statistician*, 38(4):244–247, 1984.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramigas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundations, 1991.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

J. E. Griffin and P. J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

M. Gupta and J. G. Ibrahim. Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association*, 102(479): 867–880, 2007.

C. Han and B. Carlin. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001.

C. Hans, A. Dobra, and M. West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

M. J. Heaton and J. G. Scott. Bayesian computation and the linear model. In M.-H. Chen, D. Dey, P. Mueller, D. Sun, and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis*. Springer, 2010.

A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, third edition, 1961.

M. Jerrum and A. Sinclair. The Markov Chain Monte Carlo method: an approach to approximate counting and integration. In D. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, pages 482–520, Boston, 1996. PWS Publishing.

V. E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(5):689–701, 2005.

V. E. Johnson. Properties of Bayes factors based on test statistics. *Scandinavian Journal of Statistics*, 35(2):354–368, 2008.

V. E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 72(2):143–170, 2010.

V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400, 2005.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995.

R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934, 1995.

A. Krishna, H. D. Bondell, and S. K. Ghosh. Bayesian variable selection using an adaptive powered correlation prior. *Journal of Statistical Planning and Inference*, 139(8):2665–2674, 2009.

S. Kundu and D. B. Dunson. Bayes variable selection in semiparametric linear models. *Journal of the American Statistical Association*, 2014. Epub ahead of print.

E. Ley and M. F. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674, 2009.

E. Ley and M. F. Steel. Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics*, 171(2):251–266, 2012.

F. Liang and W. H. Wong. Evolutionary Monte Carlo: applications to C_p model sampling and change point problem. *Statistica Sinica*, 10:317–342, 2000.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

B. Liquet and M. Chadeau-Hyam. *R2GUESS: Wrapper functions for GUESS*, 2014. URL `http://CRAN.R-project.org/package=R2GUESS`. R package version 1.1.

B. Liquet, M. Chadeau-Hyam, L. Bottolo, G. Campanella, and S. Richardson. *GUESS: Graphical unit evolutionary stochastic search for Bayesian model exploration*. Bayesian Integrative Genomics group, 2013. URL `http://www.bgx.org.uk/software/guess.html`.

L. Ma. Bayesian recursive variable selection. Technical report, Duke University, 2012. URL `http://stat.duke.edu/~lm186/files/ModelSel.pdf`.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89 (428):1535–1546, 1994.

D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232, 1995.

D. Madigan, A. E. Raftery, J. York, J. M. Bradshaw, and R. G. Almond. Strategies for graphical model selection. In P. Cheeseman and R. Oldford, editors, *Selecting Models from Data*, volume 89 of *Lecture Notes in Statistics*, pages 91–100. Springer, New York, 1994.

M. A. Martinez-Beneito, G. García-Donato, and D. Salmerón. A Bayesian Joinpoint regression model with an unknown number of break-points. *Annals of Applied Statistics*, 5(3):2150–2168, 2011.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman and Hall, New York, second edition, 1989.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

D. J. Nott and D. Leonte. Sampling schemes for Bayesian variable selection in generalized linear models. *Journal of Computational and Graphical Statistics*, 13(2):362–382, 2004.

I. Ntzoufras, P. Dellaportas, and J. J. Forster. Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111(1-2):165–180, 2003.

A. Raftery, J. Hoeting, C. Volinsky, I. Painter, and K. Y. Yeung. *BMA: Bayesian model averaging*, 2013. URL `http://CRAN.R-project.org/package=BMA`. R package version 3.16.2.3.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

D. Rossell, J. D. Cook, D. Telesca, and P. Roebuck. *mombf: Moment and inverse moment Bayes factors*, 2013. URL `http://CRAN.R-project.org/package=mombf`. R package version 1.5.4.

P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3):429–467, 1994.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(2):319–392, 2009.

D. Sabanés Bové and L. Held. Bayesian fractional polynomials. *Statistics and Computing*, 21(3): 309–324, 2011.

X. Sala-I-Martin, G. Doppelhofer, and R. I. Miller. Determinants of long-term growth: a Bayesian averaging of classical estimates (BACE) approach. *American Economic Review*, 94 (4):813–835, 2004.

C. Schäfer and N. Chopin. Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing*, 23(2):163–184, 2013.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619, 2010.

J. G. Scott and C. M. Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):790–808, 2008.

T. Sellke, M. J. Bayarri, and J. O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

N. Sha, M. Vannucci, M. G. Tadesse, P. J. Brown, I. Dragoni, N. Davies, T. C. Roberts, A. Contestabile, M. Salmon, C. Buckley, and F. Falciani. Bayesian Variable Selection in Multinomial Probit Models to Identify Molecular Signatures of Disease Stage. *Biometrics*, 60(3):812–819, 2004.

A. M. Strasak, N. Umlauf, R. M. Pfeiffer, and S. Lang. Comparing penalized splines and fractional polynomials for flexible modelling of the effects of continuous predictor variables. *Computational Statistics and Data Analysis*, 55(4):1540–1551, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58(1):267–288, 1996.

M. P. Wand and J. T. Ormerod. On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50(2):179–198, 2008.

A. Womack, Gopal, V., Novelo, L. L., Casella, and G. *varSelectIP: Objective Bayes model selection*, 2013. URL `http://CRAN.R-project.org/package=varSelectIP`. R package version 0.2-1.

J. York, D. Madigan, I. Heuch, and R. T. Lie. Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(2):227–242, 1995.

A. Zellner. Applications of Bayesian analysis in econometrics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):23–34, 1983.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, chapter 5, pages 233–243. North-Holland, Amsterdam, 1986.

A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, pages 585–603, Valencia, 1980. University of Valencia Press.

PAPER I

Hyper-g priors for generalized linear models

Daniel Sabanés Bové & Leonhard Held

Paper published in *Bayesian Analysis*, 2011, **6**, 387–410.

Bayesian Analysis (2011) **6**, Number 3, pp. 387–410

Hyper-g Priors for Generalized Linear Models

Daniel Sabanés Bové* and Leonhard Held†

Abstract. We develop an extension of the classical Zellner's g-prior to generalized linear models. Any continuous proper hyperprior $f(g)$ can be used, giving rise to a large class of hyper-g priors. Connections with the literature are described in detail. A fast and accurate integrated Laplace approximation of the marginal likelihood makes inference in large model spaces feasible. For posterior parameter estimation we propose an efficient and tuning-free Metropolis-Hastings sampler. The methodology is illustrated with variable selection and automatic covariate transformation in the Pima Indians diabetes data set.

Keywords: g-prior, generalized linear model, integrated Laplace approximation, variable selection, fractional polynomials

1 Introduction

Assume that we have observed n independent responses y_i coming from a generalized linear model (GLM, see McCullagh and Nelder 1989) incorporating the covariate vectors $\boldsymbol{x}_i \in \mathbb{R}^p$ via the linear predictors $\eta_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$, $i = 1, \ldots, n$. The response function (inverse link function) h transforms η_i to the mean $\mathbb{E}(y_i) = \mu_i = h(\eta_i)$, which in turn is mapped to the canonical parameter $\theta_i = (db/d\theta)^{-1}(\mu_i)$ of the exponential family. Here $db/d\theta$ is the first derivative of the function b as defined in the likelihood for $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ via

$$f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}) \propto \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right\}, \tag{1}$$

where each θ_i depends on the intercept β_0 and the vector $\boldsymbol{\beta}$ of regression coefficients as described above. Often the canonical response function $h = db/d\theta$ is used where $\theta_i = \eta_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$. The dispersions $\phi_i = \phi/w_i$ are assumed known and can incorporate weights w_i. The variance $\mathrm{Var}(y_i) = \phi_i d^2b/d\theta^2(\theta_i)$ is expressed through the variance function $v(\mu_i) = d^2b/d\theta^2((db/d\theta)^{-1}(\mu_i))$ as $\mathrm{Var}(y_i) = \phi_i v(\mu_i)$.

A Bayesian analysis starts by assigning prior distributions to the unknown model parameters β_0 and $\boldsymbol{\beta}$. However, usually there is not only uncertainty with respect to the model parameters, but also to the model itself, see e. g. Clyde and George (2004). Let γ be the model index contained in some model space Γ. Typically, the variable selection problem is considered, where $\gamma \in \{0, 1\}^m$ collects binary inclusion indicators for all m available covariates. Here we think more generally of uncertainty about the form (including the dimension p_γ) of the covariate vectors $\boldsymbol{x}_{\gamma i}$, which may also comprise

*Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland, mailto:daniel.sabanesbove@ifspm.uzh.ch

†Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland, mailto:leonhard.held@ifspm.uzh.ch

 DOI:10.1214/11-BA615

different transformations of the original variables. For example, when γ indicates a quadratic transformation of x_i, then $\boldsymbol{x}_{\gamma i} = (x_i, x_i^2)^T$. Thus, priors $f(\beta_0, \boldsymbol{\beta}_\gamma \,|\, \gamma)$ need to be assigned, for all models $\gamma \in \Gamma$. Manual elicitation of all these priors is clearly infeasible when Γ is large. In this situation priors which automatically derive from γ are attractive, and we will propose such priors in this paper. Model inference then uses the posterior model probabilities

$$f(\gamma \,|\, \boldsymbol{y}) \propto f(\boldsymbol{y} \,|\, \gamma) f(\gamma), \quad \gamma \in \Gamma, \tag{2}$$

which combine the marginal likelihood

$$f(\boldsymbol{y} \,|\, \gamma) = \int_{\mathbb{R}^{p_\gamma+1}} f(\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_\gamma, \gamma) f(\beta_0, \boldsymbol{\beta}_\gamma \,|\, \gamma)\, d\beta_0 d\boldsymbol{\beta}_\gamma \tag{3}$$

with the prior model probabilities $f(\gamma)$.

In the special case of the classical normal linear model with known error variance ϕ and $w_i \equiv 1$, the g-prior for the regression coefficients was proposed by Zellner (1986) as a "reference informative prior". It is a mean-zero normal distribution with covariance matrix $g\phi(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}$,

$$\boldsymbol{\beta}_\gamma \,|\, g, \phi \sim \mathrm{N}_{p_\gamma}\big(\mathbf{0}_{p_\gamma}, g\phi(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}\big), \tag{4}$$

and is usually combined with a locally uniform (Jeffreys) prior on β_0, assuming that the design matrix $\boldsymbol{X}_\gamma = (\boldsymbol{x}_{\gamma 1}, \ldots, \boldsymbol{x}_{\gamma n})^T$ has been centered to ensure $\boldsymbol{X}_\gamma^T \mathbf{1}_n = \mathbf{0}_{p_\gamma}$ (see Fernández et al. 2001). Often also the error variance ϕ is assumed unknown and assigned a Jeffreys prior.

The g-prior can be interpreted as the conditional posterior of $\boldsymbol{\beta}_\gamma$ given a locally uniform prior and an imaginary sample $\boldsymbol{y}_0 = \mathbf{0}_n$ from the normal linear model with design matrix $\boldsymbol{X}_\gamma$ and scaled error variance $g\phi$. This reflects the idea that after accounting for the mean level β_0 not included in the g-prior, there is no difference between observations due to the covariates in $\boldsymbol{X}_\gamma$ modelled through $\boldsymbol{\beta}_\gamma$. In addition to this nice interpretation, the g-prior has other advantages, such as invariance of the implied prior for the linear predictor under rescaling and translation of the covariates (Robert and Saleh 1991, p. 71), and automatic adaption to situations with near-collinearity between different covariates (Robert 2001, p. 193).

The hyperparameter $g > 0$ in (4) acts as an inverse relative prior sample size, hence its influence on the results is quite strong. Larger values of g lead to preference of less complex models, a phenomenon known as the Lindley-Jeffreys paradox (Lindley 1957; see also Robert et al. 2009, p. 161). Therefore, much research has been done in developing automatic specifications of g (George and Foster 2000; Hansen and Yu 2001; Fernández et al. 2001; Cui and George 2008). Moreover, a fixed g does not allow the Bayes factor of a perfectly fitting model versus the null model go to infinity (Berger and Pericchi 2001). The multivariate Cauchy priors of Zellner and Siow (1980) correspond to fully Bayesian inference with an inverse-gamma hyperprior for g. Unfortunately, the corresponding marginal likelihood $f(\boldsymbol{y} \,|\, \gamma)$ has no closed form. Therefore Liang

et al. (2008) proposed the hyper-g prior, which is a special case of the incomplete inverse-gamma prior by Cui and George (2008). These hyperpriors retain a closed form expression for $f(\boldsymbol{y} \mid \gamma)$ which is vital for efficient model inference.

In this article we develop an extension of the classical g-prior (4) to GLMs. The hyperprior on the hyperparameter g is handled in a flexible way, so that any continuous proper hyperprior $f(g)$ can be used. In Section 2, this generalized hyper-g prior is derived and connections with the literature are described. Because model inference is the main practical use of this automatic prior formulation, we will propose a fast and accurate numerical approximation of the marginal likelihood in Section 3. Section 3 also covers posterior parameter estimation with a tuning-free Markov chain Monte Carlo (MCMC) sampler. The methodology is applied to variable selection in Section 4 and to fractional polynomial modelling in Section 5. Section 6 discusses possibilities for future research.

2 The generalized hyper-g prior

Section 2.1 derives the generalized hyper-g prior, using arguments analogous to the standard g-prior. Several similar proposals can be found in the literature and are described in Section 2.2.

2.1 Prior construction

Consider the imaginary sample $\boldsymbol{y}_0 = h(0)\mathbf{1}_n$ from the GLM with design matrix $\boldsymbol{X}_\gamma$ (not including an intercept column $\mathbf{1}_n$), original weights vector $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ and scaled dispersion $g\phi$. Using an improper flat prior for the regression coefficients vector $\boldsymbol{\beta}_\gamma$, its posterior given $\boldsymbol{y}_0$ is proportional to the likelihood (1),

$$f(\boldsymbol{\beta}_\gamma \mid \boldsymbol{y}_0, g, \gamma) \propto \exp\left\{ \frac{1}{g\phi} \sum_{i=1}^{n} \Big[h(0) w_i \theta_i - w_i b(\theta_i) \Big] \right\}. \tag{5}$$

This distribution can be recognized as the Chen and Ibrahim (2003, formula 2.6) prior, although the authors have only considered the case $w_i \equiv 1$ and include the intercept β_0. Similar to their theorem 3.1, we can prove that the mode of this distribution is at $\boldsymbol{\beta}_\gamma = \mathbf{0}_{p_\gamma}$ (see the Appendix). It results from standard Bayesian asymptotic theory (e. g. Bernardo and Smith 2000, p. 287) that this distribution converges for $n \to \infty$ to the normal distribution

$$\boldsymbol{\beta}_\gamma \mid g, \gamma \sim \mathrm{N}_{p_\gamma}\big(\mathbf{0}_{p_\gamma}, g\phi c(\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma)^{-1}\big) \tag{6}$$

where $c = v(h(0)) \cdot dh/d\eta(0)^{-2}$ and $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w})$, because the inverse of the expected Fisher information $I(\boldsymbol{\beta}_\gamma)$ evaluated at the mode is $I(\mathbf{0}_{p_\gamma})^{-1} = g\phi c(\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma)^{-1}$ (cf. Chen and Ibrahim 2003, theorem 2.3).

The "generalized g-prior" (6) differs from the standard g-prior (4) only by the constant c and the weight matrix $\boldsymbol{W}$. Both are especially important in binomial regression

Family	Link	c
Gaussian	Identity	1
	(Log)	1
Poisson	Log	1
	Identity	(0)
Bernoulli	Logit	4
	Cauchit	$\pi^2/4$
	Probit	$\pi/2$
	Complementary log-log	$e-1$
Gamma	Log	1
Inverse-Gaussian	(Log)	1

Table 1: Exponential families, usual link functions and resulting factors c. Note that for the gamma and the inverse-Gaussian family, the natural links μ^{-1} and μ^{-2}, respectively, cannot be used because then $h(0) = \infty$. Parenthesized links should not be used because the uniqueness of the prior mode at $\boldsymbol{\beta}_\gamma = \mathbf{0}_{p_\gamma}$ is not sure (Wedderburn 1976). Parenthesized c's point out problems there.

when w_i is the sample size of the observed proportion, say $y_i = s_i/w_i$ if $s_i \sim \text{Bin}(w_i, \mu_i)$ is the number of successes: In Table 1 it can be seen that only for the Bernoulli family $c \neq 1$. While technically, this scaling constant could be subsumed into the hyperprior on g, it is important because it preserves the interpretation of g as the inverse relative prior sample size, i. e., the prior contains $1/g$ as much information as the data $\boldsymbol{y}$. The use of a common hyperprior $f(g)$ for different exponential families is thus simplified because g always has the same meaning. Although binomial data can always be rephrased as binary data with appropriately replicated covariate vectors and weights $w_i \equiv 1$, this is not possible for non-integer weights w_i where $\boldsymbol{W}$ is absolutely necessary. Non-integer weights are used, for example, for inverse probability weighting (Robins et al. 2000), as sampling weights for survey data (Pfeffermann 1993) and in geographically weighted regression (Brunsdon et al. 1998). Furthermore, note that the g-prior for the normal linear model with independent heteroscedastic errors $\varepsilon_i \sim \text{N}(0, \phi/w_i)$ naturally arises from (6).

Since the intercept β_0 parametrizes the average linear predictor in each model, we can use an improper flat prior $f(\beta_0) \propto 1$. Thus, our generalized g-prior does not shrink the intercept towards zero, while the mean-zero prior on the regression coefficients reflects the idea that $\boldsymbol{X}_\gamma$ has *a priori* no effect on the centered observations. The factor g is assigned a (continuous) hyperprior $f(g)$. In our approach $f(g)$ must be proper to ensure that Bayes factor comparisons with the null model, which does not include the parameter g, are valid. Apart from that, $f(g)$ can be chosen at complete liberty. As g is assigned a hyperprior, we call the resulting prior a "generalized hyper-g prior".

2.2 Comparison with the literature

An immediate question is why we do not use the exact Chen and Ibrahim (2003) prior, which is also a generalization of the standard g-prior. The main problem with this conjugate prior given in (5) is that it does not have a closed form for non-normal exponential families, where the normalizing constant of (5) is unknown. This complicates the computation of the marginal likelihood and the MCMC sampling considerably. Chen et al. (2008) propose a solution where they run an MCMC sampler on the full model, and then derive estimates for submodels. However, this approach is not applicable in problems with simultaneous variable selection and transformation as that presented in Section 5, because no full model exists in that case. Regarding the hyperparameter g, Chen and Ibrahim (2003) propose to assign it an inverse-gamma hyperprior.

Alternatively, Gupta and Ibrahim (2009) proposed the information matrix prior, which uses the expected Fisher information matrix $I(\boldsymbol{\beta}_\gamma)$ similarly to a precision matrix for a normal distribution up to a scalar variance factor g:

$$f_{GI}(\boldsymbol{\beta}_\gamma \mid g, \gamma) \propto |I(\boldsymbol{\beta}_\gamma)|^{1/2} \exp\left\{-\frac{1}{2g}\boldsymbol{\beta}_\gamma^T I(\boldsymbol{\beta}_\gamma)\boldsymbol{\beta}_\gamma\right\}. \tag{7}$$

This will only be a Gaussian distribution if the matrix $I(\boldsymbol{\beta}_\gamma)$ actually does not depend on $\boldsymbol{\beta}_\gamma$, e. g. for the normal linear model where the standard g-prior is reproduced by (7). By contrast, the precision of our generalized g-prior in (6) results from evaluating $I(\boldsymbol{\beta}_\gamma)$ at the prior mode, producing a matrix which does not depend on $\boldsymbol{\beta}_\gamma$. Gupta and Ibrahim (2009) fix the hyperparameter g at a "moderately large" value ($g \geq 1$) and do not consider inference for it.

The information matrix prior is strongly linked with the unit information prior approach of Kass and Wasserman (1995), who proposed the general idea that the amount of information in the prior on $\boldsymbol{\beta}_\gamma$ should be equal to the amount of information about it contained in one observational unit. The amount of information is measured by the (expected) Fisher information, so that the precision is chosen as $n^{-1}I(\mathbf{0}_{p_\gamma})$ in the normal prior

$$f_{KW}(\boldsymbol{\beta}_\gamma \mid g, \gamma) = \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \mid \mathbf{0}_{p_\gamma}, nI(\mathbf{0}_{p_\gamma})^{-1}\right). \tag{8}$$

This proposal is close to ours in (6), except that the hyperparameter is fixed at $g = n$. Note that Kass and Wasserman (1995) also required the nuisance parameter β_0 to be (null-)orthogonal to the parameter of interest $\boldsymbol{\beta}_\gamma$, which we ensure by centering the covariates around zero. The unit information prior was used by Ntzoufras et al. (2003) and Overstall and Forster (2010) in the GLM context.

Hansen and Yu (2003, p. 156) also use the expected Fisher information, but evaluate it at the maximum likelihood (ML) estimate $\widehat{\boldsymbol{\beta}}_\gamma$ to obtain a prior precision matrix:

$$f_{HY}(\boldsymbol{\beta}_\gamma \mid g, \gamma) = \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \mid \mathbf{0}_{p_\gamma}, gI(\widehat{\boldsymbol{\beta}}_\gamma)^{-1}\right). \tag{9}$$

Hansen and Yu find the dependence of their prior on the data $\boldsymbol{y}$ "hard to accept", although it can be interpreted as an empirical Bayes approach. Also in this flavour, the

authors maximize a cost-modified (approximate) likelihood of g in order to eliminate g. Subsequent model selection is then based on this function value ("minimum description length").

Instead of using the *expected* Fisher information matrix $I(\boldsymbol{\beta}_\gamma)$, Wang and George (2007) use the *observed* Fisher information matrix $J(\boldsymbol{\beta}_\gamma)$. While for canonical response functions the equality $I(\boldsymbol{\beta}_\gamma) = J(\boldsymbol{\beta}_\gamma)$ holds, in general $J(\boldsymbol{\beta}_\gamma)$ is different and depends on the observed response vector. Wang and George (2007) evaluate the observed Fisher information at the original response $\boldsymbol{y}$ and the ML estimate $\widehat{\boldsymbol{\beta}}_\gamma$ to obtain the correlation structure of the normal distribution:

$$f_{WG}(\boldsymbol{\beta}_\gamma \,|\, g, \gamma) = \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \,|\, \mathbf{0}_{p_\gamma}, gJ(\widehat{\boldsymbol{\beta}}_\gamma)^{-1}\right). \tag{10}$$

By comparison, our generalized g-prior (6) does not use the original data $\boldsymbol{y}$, but only the design matrix $\boldsymbol{X}_\gamma$. Analogously to Hansen and Yu (2003), Wang and George (2007) select model-specific values for g by maximizing $f(\boldsymbol{y} \,|\, g, \gamma)$, but they also consider fully Bayesian inference for g with flat or truncated-gamma hyperpriors on $1/(g+1)$.

Marin and Robert (2007, p. 101) avoid the use of a Fisher information matrix altogether when they propose the "non-informative g-prior"

$$f_{MR}(\boldsymbol{\beta}_{0\gamma} \,|\, g, \gamma) = \mathrm{N}_{p_\gamma+1}\left(\boldsymbol{\beta}_{0\gamma} \,|\, \mathbf{0}_{p_\gamma+1}, g(\boldsymbol{X}_{0\gamma}^T\boldsymbol{X}_{0\gamma})^{-1}\right) \tag{11}$$

for binary regression with probit and logit link functions, where $\boldsymbol{\beta}_{0\gamma} = (\beta_0, \boldsymbol{\beta}_\gamma^T)^T$ denotes the vector of all coefficients with corresponding full design matrix $\boldsymbol{X}_{0\gamma} = (\mathbf{1}_n, \boldsymbol{X}_\gamma)$. Thus, the intercept β_0 is included in the g-prior. Note that also Gupta and Ibrahim (2009), Hansen and Yu (2003) and Wang and George (2007) originally do not separate the intercept from the other regression coefficients. When $\boldsymbol{X}_\gamma$ is not centered, the intercept is then *a priori* correlated with the other coefficients. In addition, it is also shrunk to its prior mean, not necessarily a desired feature in applications. Marin and Robert (2007) are able to assign g an improper hyperprior, $f(g) \propto g^{-3/4}$, which can be regarded as a degenerate inverse-gamma distribution with shape $-1/4$ and scale 0, because the hyperparameter g is also included in the null model with intercept only.

3 Implementation

In Section 3.1 we propose an accurate numerical approximation of the marginal likelihood under the generalized hyper-g prior. Given a specific model, we can sample from the posterior using a tuning-free Metropolis-Hastings scheme described in Section 3.2. In Section 3.3 we investigate the performance of the numerical and an MCMC marginal likelihood approximation in the conjugate setup, where exact values are known.

3.1 Marginal likelihood computation

Under the generalized hyper-g prior, the marginal likelihood (3) of the GLM γ is

$$\begin{aligned} f(\boldsymbol{y} \mid \gamma) &= \int_{\mathbb{R}^{p_\gamma+1}} f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_\gamma, \gamma) \int_{\mathbb{R}_+} f(\boldsymbol{\beta}_\gamma \mid g, \gamma) f(g)\, dg\, d\beta_0 d\boldsymbol{\beta}_\gamma \\ &= \int_{\mathbb{R}_+} f(\boldsymbol{y} \mid g, \gamma) f(g)\, dg, \end{aligned} \tag{12}$$

where

$$f(\boldsymbol{y} \mid g, \gamma) = \int_{\mathbb{R}^{p_\gamma+1}} f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_\gamma, \gamma) f(\boldsymbol{\beta}_\gamma \mid g, \gamma)\, d\beta_0 d\boldsymbol{\beta}_\gamma \tag{13}$$

is the likelihood of g. Note that both (12) and (13) are only defined up to a constant which we have fixed at unity, as we use the improper prior $f(\beta_0) \propto 1$. In general, no closed form expressions are available. The obvious exception is the special case of a Gaussian likelihood, which was mentioned in Section 1 and will be referred to again in Section 3.3. Therefore, in order to be able to efficiently explore a large model space Γ, we need to develop a fast but accurate numerical approximation to the marginal likelihood. This will be a two-step procedure: The likelihood of g in (13) is computed by a Laplace approximation. Plugging this into (12), the hyperparameter g will be integrated out with respect to its prior by numerical integration. Together, this is an integrated Laplace approximation (ILA), which was proposed more generally by Rue et al. (2009).

The Laplace approximation (Lindley 1980; Tierney and Kadane 1986) of (13) is

$$\begin{aligned} f(\boldsymbol{y} \mid g, \gamma) &\approx \frac{f(\boldsymbol{y} \mid \boldsymbol{\beta}^*_{0\gamma}, \gamma) f(\boldsymbol{\beta}^*_{0\gamma} \mid g, \gamma)}{\tilde{f}(\boldsymbol{\beta}^*_{0\gamma} \mid \boldsymbol{y}, g, \gamma)} \\ &= f(\boldsymbol{y} \mid \boldsymbol{\beta}^*_{0\gamma}, \gamma) (2\pi)^{(p+1)/2} \left|\boldsymbol{R}^*_{0\gamma}\right|^{-1/2} \\ &\quad \times (2\pi g\phi c)^{-p/2} \left|\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma\right|^{1/2} \exp\left\{-\frac{1}{2}(g\phi c)^{-1} \boldsymbol{\beta}^{*T}_\gamma \boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma \boldsymbol{\beta}^*_\gamma\right\} \end{aligned} \tag{14}$$

when $\tilde{f}(\boldsymbol{\beta}_{0\gamma} \mid \boldsymbol{y}, g, \gamma)$ is the Gaussian approximation of the conditional coefficients posterior with mean vector $\boldsymbol{\beta}^*_{0\gamma}$ and precision matrix $\boldsymbol{R}^*_{0\gamma}$. Since the conditional coefficients prior can be seen to have a normal kernel $f(\boldsymbol{\beta}_{0\gamma} \mid g, \gamma) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T_{0\gamma} \boldsymbol{R}_{0\gamma} \boldsymbol{\beta}_{0\gamma}\right\}$ with (singular) precision

$$\boldsymbol{R}_{0\gamma} = \operatorname{diag}\left\{0, (g\phi c)^{-1} \boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma\right\}, \tag{15}$$

the Bayesian iterative weighted least squares (IWLS) algorithm (West 1985; Gamerman 1997) can be used to compute the moments of the Gaussian approximation. Note that this is different and potentially more accurate than the approach by Rue et al. (2009, p. 327) who preserve the sparsity of the prior precision $\boldsymbol{R}_{0\gamma}$ in the resulting posterior precision $\boldsymbol{R}^*_{0\gamma}$. The accuracy of the Laplace approximation (14) can be even further improved by including higher-order terms of the underlying Taylor expansion. For canonical response functions, Raudenbush et al. (2000) derived a convenient correction

factor corresponding to a sixth-order Laplace approximation. In the applications of Sections 4 and 5, we have used this correction (see the Appendix for details), which clearly improved the ILA while requiring only slightly more computation time.

The one-dimensional integration in (12) is performed on the log-scale over $z = \log(g)$ using Gauss-Hermite quadrature. First, we find the (approximate) posterior mode z^* and variance σ^{*2} of z using its unnormalized (approximate) posterior density

$$\tilde{f}(z, \boldsymbol{y} \,|\, \gamma) = \tilde{f}(\boldsymbol{y} \,|\, z, \gamma) f(z). \tag{16}$$

The mode z^* is numerically determined by the `optimize` routine in `R` (R Development Core Team 2010; Brent 1973). The variance σ^{*2} can be computed as the negative inverse second derivative of the log posterior at z^* by numerical differentiation (routine `dfridr` from Press et al. 2007, p. 231). Second, we apply the Gauss-Hermite quadrature (Naylor and Smith 1982)

$$f(\boldsymbol{y} \,|\, \gamma) \approx \sum_{j=1}^{N} m_j \tilde{f}(z_j, \boldsymbol{y} \,|\, \gamma), \tag{17}$$

where the actual weights $m_j = \omega_j \exp(t_j^2)\sqrt{2}\sigma^*$ and nodes $z_j = z^* + \sqrt{2}\sigma^* t_j$ depend on z^*, σ^* as well as original weights ω_j and nodes t_j, $j = 1, \ldots, N$. These can be obtained from the Golub and Welsch (1969) algorithm, which is implemented in the `R`-function `gauss.quad` (Smyth et al. 2010). $N = 20$ seems to be sufficient, given that this includes nodes in a range of about seven standard deviations around z^* (as then $\sqrt{2}t_{20} \approx 7.6$). Note that the Gauss-Hermite approximation in (17) is exact if $\tilde{f}(z, \boldsymbol{y} \,|\, \gamma)$ is the product of $\mathrm{N}(z \,|\, z^*, \sigma^{*2})$ and a polynomial of at most order $2N - 1$.

3.2 Metropolis-Hastings sampler

Given a model $\gamma \in \Gamma$ we would like to sample from the joint posterior of the model-specific parameters $\boldsymbol{\theta}_\gamma = (\boldsymbol{\beta}_{0\gamma}^T, z)^T$. To this end, we propose a tuning-free Metropolis-Hastings (MH) sampling scheme with proposal kernel

$$q(\boldsymbol{\theta}'_\gamma \,|\, \boldsymbol{\theta}_\gamma) = q(\boldsymbol{\beta}'_{0\gamma} \,|\, z', \boldsymbol{\beta}_{0\gamma}) q(z') \tag{18}$$

for the proposal $\boldsymbol{\theta}'_\gamma$ given the current sample $\boldsymbol{\theta}_\gamma$. The independence proposal density $q(z)$ is constructed by first linearly interpolating pairs $\big(z_j, \tilde{f}(z_j, \boldsymbol{y} \,|\, \gamma)\big)$ and second normalizing this function to unity integral, $\int_{\min z_j}^{\max z_j} q(z)\, dz = 1$. Note that many pairs are already available from the optimization and integration of (16) in the marginal likelihood computation, and finer approximations can be obtained by incorporating suitable additional grid points z_j. Thus, $q(z)$ is close to the posterior density $f(z \,|\, \boldsymbol{y}, \gamma)$, suggesting high acceptance rates of the sampler. Also, generating random variates from $q(z)$ using inverse sampling is straightforward as the corresponding cumulative distribution function is piecewise quadratic.

For the coefficients, $q(\boldsymbol{\beta}'_{0\gamma} \,|\, z', \boldsymbol{\beta}_{0\gamma})$ is a Gaussian proposal density: Starting from the current vector $\boldsymbol{\beta}_{0\gamma}$ and the proposed prior covariance factor $g' = \exp(z')$, a single step

of the Bayesian IWLS is made, resulting in the mean vector and the precision matrix of the proposal (Gamerman 1997). In order to compute the acceptance probability of the move from $\boldsymbol{\theta}_\gamma$ to $\boldsymbol{\theta}'_\gamma$,

$$\alpha(\boldsymbol{\theta}'_\gamma \,|\, \boldsymbol{\theta}_\gamma) = 1 \wedge \frac{f(\boldsymbol{y} \,|\, \boldsymbol{\beta}'_{0\gamma}, \gamma) f(\boldsymbol{\theta}'_\gamma \,|\, \gamma)}{f(\boldsymbol{y} \,|\, \boldsymbol{\beta}_{0\gamma}, \gamma) f(\boldsymbol{\theta}_\gamma \,|\, \gamma)} \cdot \frac{q(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{\theta}'_\gamma)}{q(\boldsymbol{\theta}'_\gamma \,|\, \boldsymbol{\theta}_\gamma)}, \tag{19}$$

note that the prior contributions have the form $f(\boldsymbol{\theta}_\gamma \,|\, \gamma) = f(\boldsymbol{\beta}_\gamma \,|\, g, \gamma) f(g) g$, the last factor g being due to the change of variable $z = \log(g)$ in the proposal parametrization. For the reverse proposal kernel value $q(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{\theta}'_\gamma)$, another IWLS step starting from the proposed vector $\boldsymbol{\beta}'_{0\gamma}$ and the current factor $g = \exp(z)$ is necessary.

The MH sampler can also be used to compute an MCMC estimate of the marginal likelihood $f(\boldsymbol{y} \,|\, \gamma)$, providing an independent check of the numerical estimate presented in Section 3.1. We will use the method by Chib and Jeliazkov (2001, section 2.1), which was competitive in a review by Han and Carlin (2001) and is still a benchmark for new developments (see e. g. Nott et al. 2008). The estimate is based on the basic identity

$$f(\boldsymbol{y} \,|\, \gamma) = \frac{f(\boldsymbol{y} \,|\, \boldsymbol{\theta}^*_\gamma, \gamma) f(\boldsymbol{\theta}^*_\gamma \,|\, \gamma)}{f(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{y}, \gamma)}, \tag{20}$$

which holds for any $\boldsymbol{\theta}^*_\gamma$. Chib and Jeliazkov (2001) recommend to select $\boldsymbol{\theta}^*_\gamma$ close to the mode of $f(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{y}, \gamma)$. Detailed balance of the Markov chain ensures that the unknown posterior ordinate can be estimated by

$$f(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{y}, \gamma) \approx \frac{\sum_{j=1}^{B} \alpha(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{\theta}^{(j)}_\gamma) q(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{\theta}^{(j)}_\gamma)}{\sum_{k=1}^{B} \alpha(\boldsymbol{\theta}^{(k)}_\gamma \,|\, \boldsymbol{\theta}^*_\gamma)}, \tag{21}$$

where the $\boldsymbol{\theta}^{(j)}_\gamma$ are the posterior samples and the $\boldsymbol{\theta}^{(k)}_\gamma$ are iid draws from the proposal distribution $q(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{\theta}^*_\gamma)$. Since each acceptance probability in (21) requires two additional IWLS steps, $4B$ additional IWLS steps are required if B posterior samples are used.

3.3 Performance in the conjugate case

To investigate the performance of the proposed algorithms, we consider the special case of normal linear regression with fixed error variance ϕ. Using the g-prior (4), the conditional coefficients posterior is Gaussian,

$$f(\boldsymbol{\beta}_{0\gamma} \,|\, \boldsymbol{y}, g, \gamma) = \mathrm{N}\left(\beta_0 \,|\, \bar{y}, \phi/n\right) \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \,|\, g(g+1)^{-1} \widehat{\boldsymbol{\beta}}_\gamma,\ g(g+1)^{-1} \phi (\boldsymbol{X}^T_\gamma \boldsymbol{X}_\gamma)^{-1}\right), \tag{22}$$

where the ordinary least squares estimate $\widehat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}^T_\gamma \boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}^T_\gamma \boldsymbol{y}$ is shrunk by the factor $g/(g+1)$. Thus, the Laplace approximation (14) of the likelihood of g is exact and given by

$$f(\boldsymbol{y} \,|\, g, \gamma) = (g+1)^{-p_\gamma/2} \exp\left\{(g+1)^{-1} \left[-\frac{SSR_\gamma}{2\phi}\right]\right\} \cdot \exp\left\{-\frac{SSE_\gamma}{2\phi}\right\}, \tag{23}$$

where SSE_γ and SSR_γ are the error and regression sums of squares, respectively. From the form of (23) we see that an inverse-gamma hyperprior $\mathrm{IG}(a, b)$ on $g + 1$ will be conjugate to this likelihood. Since $g > 0$ must be ensured, this distribution must be truncated to $(1, \infty)$, yielding the incomplete inverse-gamma prior (Cui and George 2008, p. 891)

$$f(g) = M(a, b)(g + 1)^{-(a+1)} \exp\{-b/(g + 1)\} \tag{24}$$

with normalising constant

$$M(a, b) = \frac{b^a}{\int_0^b t^{a-1} \exp(-t)\, dt} \tag{25}$$

and corresponding marginal likelihood

$$f(\boldsymbol{y} \,|\, \gamma) = \frac{M(a, b)}{M(a_\gamma, b_\gamma)} \exp\left\{-\frac{SSE_\gamma}{2\phi}\right\}, \tag{26}$$

where the updated parameters $a_\gamma = a + p_\gamma/2$ and $b_\gamma = SSR_\gamma/(2\phi) + b$ determine the posterior of g in model γ.

For illustration, we consider the ozone data introduced by Breiman and Friedman (1985) in the notation of Sabanés Bové and Held (2010), where $n = 330$. Deciding whether to include each of the nine meteorological covariates z_0 and $z_4, \ldots, z_{11}$ in the linear regression of the daily maximum ozone concentration y yields a model space Γ of size $2^9 = 512$. For all $\gamma \in \Gamma$, the ILA (17) and the MCMC estimate (20) of the exact marginal likelihood value (26) were computed fixing the variance at $\phi = 19.75$ (the estimate in the full ordinary linear model) and using the hyperprior parameters $a = 0.01, b = 0.01$. Figure 1 shows that the errors of the ILA and the MCMC estimates are very small here compared to the absolute true values.

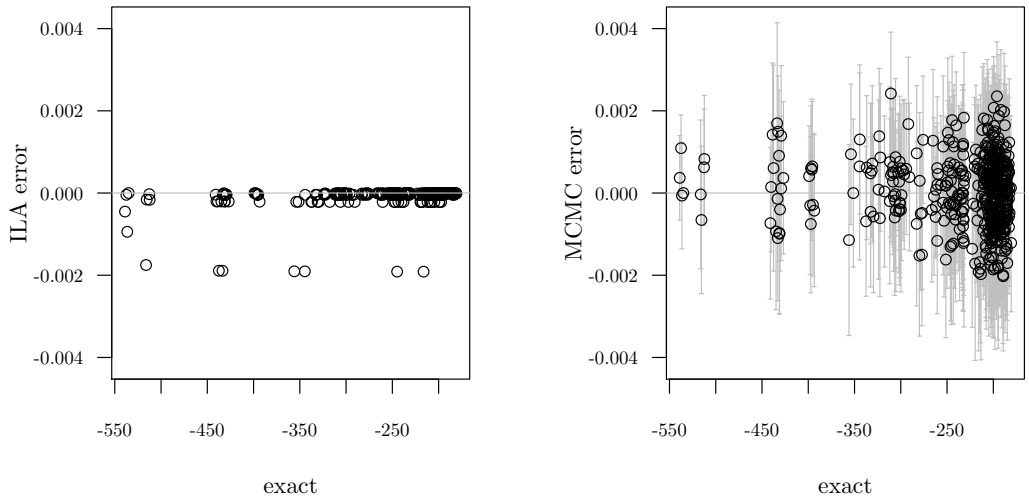
For all models, the acceptance rates of the MH algorithm were above 97%. Figure 2 shows that even for the model with the lowest acceptance rate, the true posterior density of $z = \log(g)$ is very close to its ILA estimate $q(z)$. This explains the almost perfect acceptance rates of the MH scheme.

4 Variable selection

We illustrate the methodology for non-normal data with the Pima Indians diabetes data set (Frank and Asuncion 2010; Ripley 1996), which contains $n = 532$ complete records on diabetes presence and $m = 7$ associated covariates described in Table 2. First, we restrict ourselves to variable selection in the logistic regression model, yielding a model space Γ of size $2^7 = 128$. In Section 5, we will also consider power transformations of the covariates.

Three different hyperprior distributions for the covariance factor g are compared for a fully Bayesian analysis:

F1 $f(g) = \mathrm{IG}(g \,|\, 1/2, n/2)$, corresponding to the Zellner and Siow (1980) approach;



(a) Errors of the ILA estimates.

(b) Errors of the MCMC estimates. The vertical bars represent 95% MCMC confidence intervals (coverage is 95.1% here).

Figure 1: Errors of the ILA and the MCMC estimates (y-axes) compared to the exact log marginal likelihood values (x-axes) for all 512 models. The MCMC estimates are based on $B = 4500$ samples which were saved after burn-ins of length 1000 (every 2nd iteration). Note that the log marginal likelihood values include the additional additive term $\log \sqrt{2\pi\phi/n}$ compared to (26).

Variable	Description
y	Signs of diabetes according to WHO criteria (Yes = 1, No = 0)
x_1	Number of pregnancies
x_2	Plasma glucose concentration in an oral glucose tolerance test [mg/dl]
x_3	Diastolic blood pressure [mm Hg]
x_4	Triceps skin fold thickness [mm]
x_5	Body mass index (BMI) [kg/m^2]
x_6	Diabetes pedigree function
x_7	Age [years]

Table 2: Description of the variables in the Pima Indians diabetes data set.

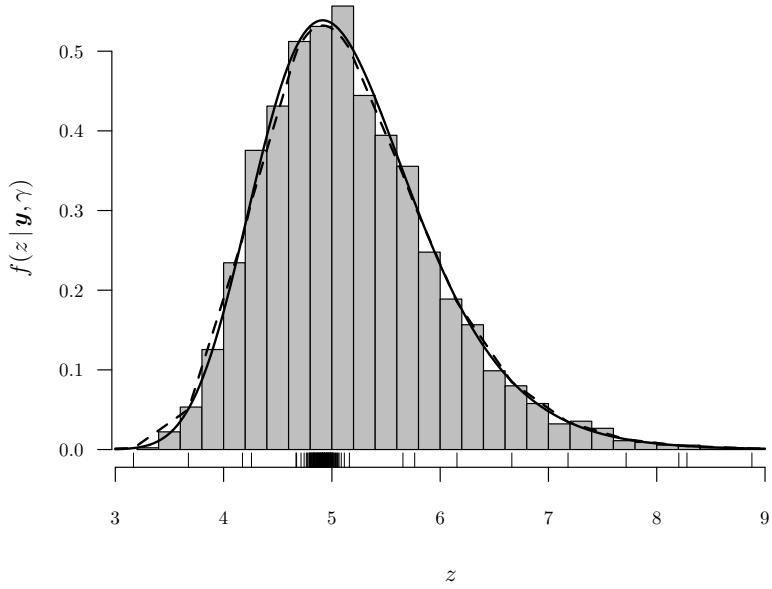


Figure 2: True posterior density of z (solid line) compared with the ILA (dashed line) and MCMC (histogram) estimates. Small ticks above the horizontal axis indicate where nodes z_j for the construction of the ILA estimate $q(z)$ were located (cf. Section 3.2).

F2 $f(g) = 1/n(1 + g/n)^{-2}$, corresponding to the hyper-g/n prior (Liang et al. 2008, p. 416);

F3 $f(g) = \mathrm{IG}(g \mid 0.001, 0.001)$, which is a standard choice for variance parameters.

We also consider model-specific empirical Bayes estimation of g using the likelihood of g in (13), abbreviating this approach as **EB**. Moreover, the standard criteria **AIC** and **BIC** are computed for each model. We use the prior model probabilities

$$f(\gamma) = \frac{1}{m+1}\binom{m}{p_\gamma}^{-1} \tag{27}$$

for an appropriate multiplicity adjustment (George and McCulloch 1993; Scott and Berger 2010). Posterior model probabilities then follow from (2), where for EB the maximized likelihood of g in (13) and for BIC the approximation $\exp(-1/2\,\mathrm{BIC})$ (e. g. Kass and Raftery 1995) is used instead of $f(\boldsymbol{y} \mid \gamma)$. Similar model weights proportional to $\exp(-1/2\,\mathrm{AIC})$ can also be calculated for AIC as proposed by Buckland et al. (1997).

In Table 3, the resulting posterior probabilities and AIC weights for variable inclusion are shown. All methods clearly select x_1, x_2, x_5 and x_6. The corresponding model is the *maximum a posteriori* (MAP) model in F1, F2, F3 and BIC, while for EB and AIC also x_7 is included in the top model. This covariate would be included as well in the median probability model (Barbieri and Berger 2004) for all methods except BIC.

	F1	F2	F3	EB	AIC	BIC
x_1	0.961	0.965	0.968	0.970	0.972	0.946
x_2	1.000	1.000	1.000	1.000	1.000	1.000
x_3	0.252	0.309	0.353	0.384	0.309	0.100
x_4	0.248	0.303	0.346	0.376	0.296	0.103
x_5	0.998	0.998	0.998	0.998	0.998	0.997
x_6	0.994	0.995	0.996	0.996	0.998	0.987
x_7	0.528	0.586	0.629	0.659	0.670	0.334

Table 3: Posterior probabilities and AIC weights for variable inclusion in the Pima Indians diabetes data.

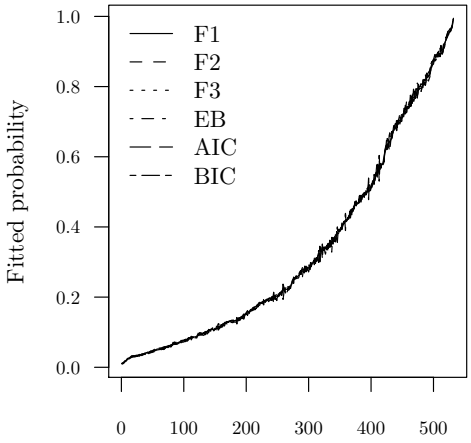
For x_3 and x_4, the evidence for inclusion is consistently weak. For comparison, Holmes and Held (2006) used vague iid normal priors for all coefficients and a flat model prior $f(\gamma) = 2^{-7}$, obtaining clear evidence for inclusion of the MAP covariates.

It is interesting that the inclusion probabilities under F1, F2 and F3 are qualitatively similar. The reason could be that the sample size is relatively large in this example, reducing the importance of the hyperprior specification for g. For EB, most inclusion probabilities are even higher than for F3. The AIC weights are more similar to F2 probabilities (except for x_7). The BIC based probabilities are mostly lower, and close to the (not shown) probabilities under F1 when a flat model prior is used.

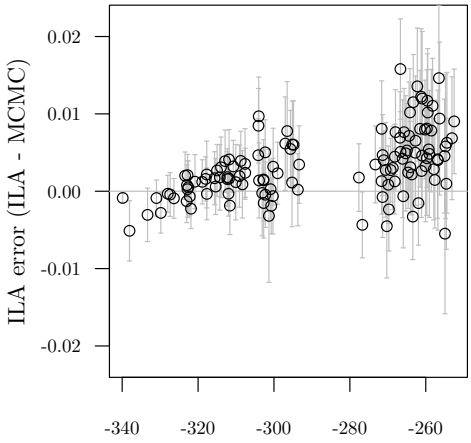
While the posterior inclusion probabilities are visibly different for the six approaches, the model-averaged fits to the data are very close, as shown in Figure 3a. In parallel to sampling the parameters leading to these fitted probabilities for F1, F2, F3 and EB, we also estimated the marginal likelihood by MCMC. The resulting MCMC estimates were close to the ILA estimates, comparison plots looking like Figure 3b for F3. Note that the coverage of the MCMC confidence intervals is lower than in Figure 1b, because the ILA approximations are not exact.

5 Fractional polynomials

Fractional polynomials (FPs) are used for systematic power transformations of the covariates $x_1, \ldots, x_m$ (Royston and Altman 1994). They widen the class of ordinary polynomials insofar as the powers are taken from the fixed set $\{-2, -1, -1/2, 0, 1/2, 1, 2, 3\}$, which also contains square roots, reciprocals and the logarithm by the Box and Tidwell (1962) convention $x^0 \equiv \log(x)$. For each covariate x_k, at most two powers are chosen and collected in the tuple $\boldsymbol{p}_k$, while the corresponding coefficients are collected in the vector $\boldsymbol{\alpha}_k$, determining the FP transform $x_k^{\boldsymbol{p}_k}\boldsymbol{\alpha}_k$. The special case $p_{k1} = p_{k2}$ is handled by multiplication with the logarithm, e. g. $x_k^{(2,2)} = \left(x_k^2, x_k^2 \log(x_k)\right)$. Variable selection is embedded in this framework, because x_k is not included in the model if $\boldsymbol{p}_k = \emptyset$. Each model is thus uniquely identified by $\gamma = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m)$, the covariate vectors are $\boldsymbol{x}_{\gamma i} = (x_{1i}^{\boldsymbol{p}_1}, \ldots, x_{mi}^{\boldsymbol{p}_m})^T$ and the vector of regression coefficients is $\boldsymbol{\beta}_\gamma = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_m^T)^T$.



(a) Model-averaged fitted probabilities.



(b) Errors of the ILA estimates with respect to the MCMC estimates of the log marginal likelihood under F3, for all 128 models. The MCMC estimates are based on (at least) $B = 5000$ samples which were saved after burn-ins of length 1000 (every 2nd iteration). The vertical bars represent 95% MCMC confidence intervals (coverage is 72.7% here).

Figure 3: Results in the Pima Indians variable selection example.

Sabanés Bové and Held (2010) implemented Bayesian model selection for normal linear FP models, and more details on FPs can be found in references therein.

The model space Γ comprises 45^m models, and thus the use of an automatic prior for the parameter $\boldsymbol{\beta}_\gamma$, conditional on the model γ, is very attractive. The generalized g-prior (6) is automatic and only depends on the global hyperparameter g. We will again compare the three fully Bayesian approaches (F1, F2, F3) with the empirical Bayes procedure (EB) which were introduced in Section 4 and avoid manual specification of g. The prior model probabilities $f(\gamma) = \prod_{k=1}^{m} f(\boldsymbol{p}_k)$ depend on the prior FP transformation probabilities

$$f(\boldsymbol{p}_k) = \frac{1}{3}\binom{7 + |\boldsymbol{p}_k|}{|\boldsymbol{p}_k|}^{-1} \tag{28}$$

which have the same form as (27): each degree $|\boldsymbol{p}_k| \in \{0, 1, 2\}$ is equally probable, and all tuples $\boldsymbol{p}_k$ of the same degree are equally probable. This implements Jeffreys's "simplicity postulate" that simpler models must have greater prior probability than more complex models (Jeffreys 1961, section 1.6), and indeed the null model has the largest prior probability 3^{-m}.

For the Pima Indinas diabetes data the model space Γ has size $45^7 \approx 3.7 \cdot 10^{11}$, rendering an exhaustive evaluation of all models $\gamma \in \Gamma$ infeasible. Therefore we use an MCMC model composition (Madigan and York 1995) approach: Starting from the null model, we move through Γ by successive slight modifications of the configuration γ. The modifications are accepted with MH acceptance probabilities, which ensures that models with higher posterior probability are more likely to be visited; see Sabanés Bové and Held (2010) for details. For all four approaches (F1, F2, F3 and EB), we ran this model sampler for one million iterations. To get an idea of the computational complexity, note that on average 10.8 (F2) and 22.1 (EB) models could be evaluated per second (on 2.8 GHz CPUs). All computations have been implemented in an `R`-package including an efficient `C++` core for the MCMC parts, which is available from the first author.

For all four approaches Table 4 shows clear evidence for inclusion of the covariates x_2, x_5, x_6 and x_7 with posterior inclusion probabilities over 99%, while the other three covariates have inclusion probabilities below 15%. In comparison with the variable inclusion results for the untransformed covariates in Table 3, it is interesting that x_1 is no longer important when FP transformations are considered, while x_7 is much more important.

In addition to examining the marginal inclusion probabilities, it is necessary to look at the transformations of the covariates. Since all four approaches produce similar variable inclusion probabilities and also share the MAP model $\boldsymbol{x}_i = (x_{2i}, x_{5i}^{-2}, x_{6i}^{-1/2}, x_{7i}^{-2})^T$, we only look at the F1 approach (the three others give very similar results). In order to account for model uncertainty, it is best to look at model-averaged estimates of variable transformations, conditional on variable inclusion. To this end we varied the transformation of one of the covariates x_2, x_5, x_6, x_7 while fixing the others at their MAP configuration. Averaging over the 44 models each results in the effect estimates shown in Figure 4. Plasma glucose concentration (x_2) seems to have a strong positive linear

	F1	F2	F3	EB
x_1	0.119	0.125	0.135	0.144
x_2	1.000	1.000	1.000	1.000
x_3	0.050	0.052	0.054	0.054
x_4	0.032	0.033	0.033	0.035
x_5	0.999	0.999	0.999	0.999
x_6	0.992	0.993	0.993	0.994
x_7	0.999	0.999	0.999	0.999

Table 4: Posterior probabilities for variable inclusion in the Pima Indians diabetes data when FP transformations are considered. The probabilities are based on 671 525 (F1), 719 929 (F2), 758 616 (F3), and 777 531 (EB) visited models.

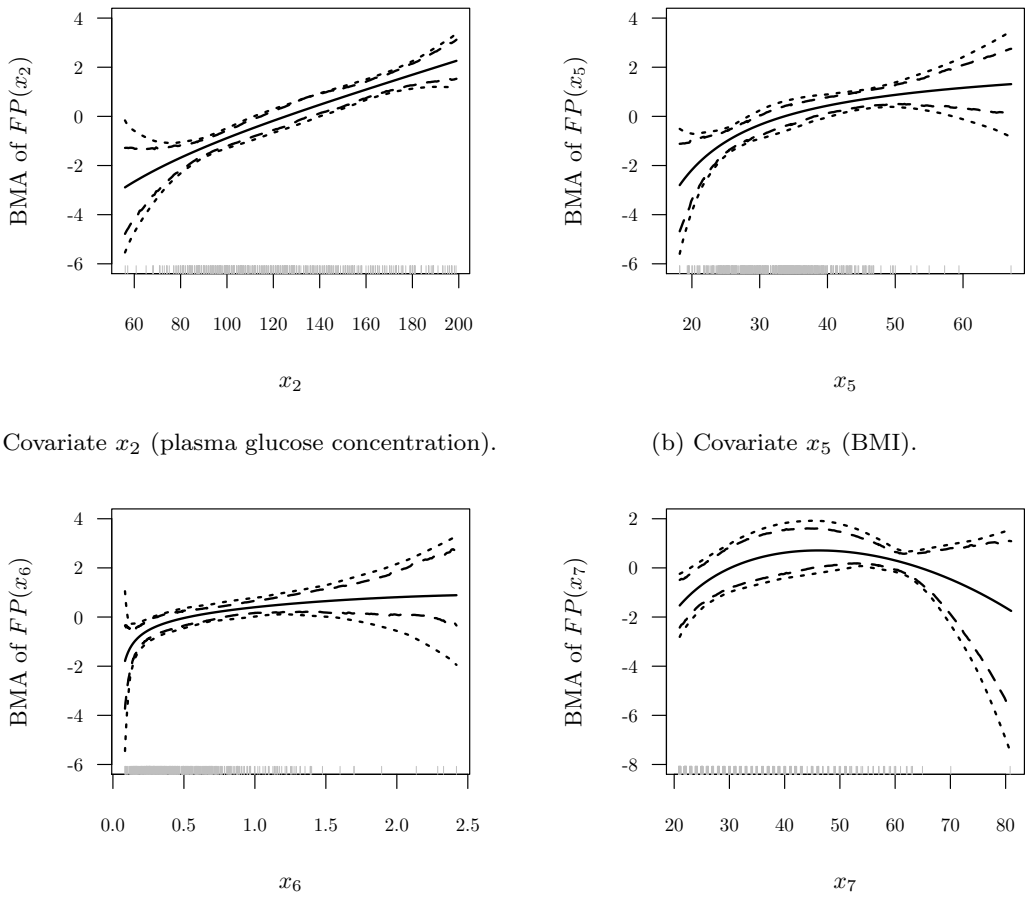
association with diabetes log-odds, while the estimated positive effect of BMI (x_5) is levelling off non-linearly for (rare) high values and is weaker overall. Even smaller is the estimated positive effect of diabetes pedigree function (x_6) with the largest increase in diabetes risk between $x_6 = 0.1$ and $x_6 = 0.5$. The estimated association of age (x_7) is clearly non-linear, with higher diabetes risk for middle-aged participants. These results are qualitatively similar to those obtained by Cottet et al. (2008, p. 665) for a larger subset of the original Pima Indians diabetes data set.

The marginal posterior distributions for the covariance factor g differ slightly between the three hyperprior choices F1, F2 and F3. Averaging over the best 1000 models in terms of posterior probability which have been visited by the model sampler, we get the histograms for $z = \log(g)$ in Figure 5. The corresponding posterior means $\mathbb{E}(g \,|\, \boldsymbol{y})$ decrease from 282.5 for F1, 219.2 for F2 to 179.1 for F3, and this trend is also visible in the histograms. The results suggest a stronger prior shrinkage of the regression coefficients than that proposed by the unit information prior's fixed value $g = n = 532$ (cf. Section 2.2), as $\mathbb{P}(g < n \,|\, \boldsymbol{y})$ ranges from 90.9% for F1 to 95.7% for F3.

6 Discussion

In this article, we presented a generalization of the g-prior to GLMs, which can be interpreted analogously to the classical g-prior for normal linear models. In our implementation, the shrinkage-controlling hyperparameter g can be assigned any hyperprior, thus giving rise to a large class of generalized hyper-g priors. For mixtures of classical g-priors, Liang et al. (2008) investigate theoretical model selection and prediction consistency properties. It would be desirable to also investigate such properties for our generalized hyper-g prior class. However, as fewer closed form expressions are available, derivation of comparable proofs will be more difficult in the GLM family.

Another important area of future research is the thorough comparison of the generalized hyper-g prior with the other approaches in the literature summarized in Section 2.2. For example, exhaustive simulation studies could shed light on different performances of



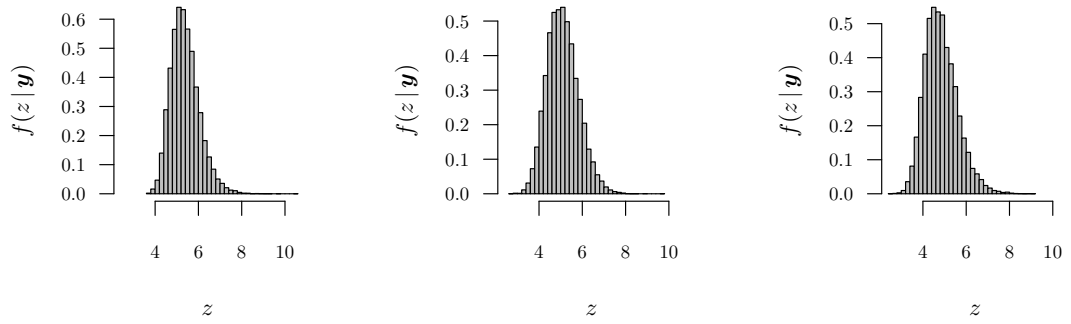
(a) Covariate x_2 (plasma glucose concentration).

(b) Covariate x_5 (BMI).

(c) Covariate x_6 (diabetes pedigree function).

(d) Covariate x_7 (age).

Figure 4: Model-averaged FP transformations of selected Pima Indians covariates under hyperprior F1. Means (solid lines), pointwise (dashed lines) as well as simultaneous (dotted lines) 95% credible intervals are given. Small ticks above the x-axes indicate data locations.



(a) Posterior under hyperprior F1.

(b) Posterior under hyperprior F2.

(c) Posterior under hyperprior F3.

Figure 5: Comparison of marginal posteriors for $z = \log(g)$ under hyperpriors F1, F2 and F3. The histograms are based on the model average over the respective 1000 models with highest posterior probability visited by the model samplers.

the priors in variable selection. Perhaps also theoretical results can be derived to explain the different properties of the approaches. An advantage of our approach is that we allow arbitrary hyperpriors for g while still providing a fast and accurate deterministic approximation to the marginal likelihood.

Bayesian model selection for FPs in GLMs was in fact the motivating application for this work. With huge model spaces to explore, the accurate numerical marginal likelihood approximation is vital for this and similar typical applications of the generalized hyper-g prior. Alternative MCMC estimates of the marginal likelihood were used to demonstrate the very good accuracy of the ILA estimates. Yet, MCMC would not be suited for replacing the deterministic ILA approach in the stochastic model search, because the computation is slower by orders of magnitude and would require careful automatic monitoring of convergence. Of course, the deterministic marginal likelihood approximation could be used for any type of stochastic model search, such as those recently proposed by Hans et al. (2007) and Dobra (2009).

Finally, we note that the classical g-prior has recently been extended in other directions as well. In the context of supervised machine learning, Zhang et al. (2009) replace $\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma$ by a (possibly singular) kernel matrix $\boldsymbol{K}_\gamma$ and prove consistency properties for the normal linear model. Maruyama and George (2010) remove the restriction of $p_\gamma \leq n-1$ for normal linear models by working with the singular value decomposition (SVD) of the design matrix $\boldsymbol{X}_\gamma$. A similar extension is the "generalized singular g-prior" defined by West (2003) in the factor regression context. Along these lines, our generalized hyper-g prior could also be extended to the $p_\gamma > n$ case via the SVD $\boldsymbol{W}^{1/2}\boldsymbol{X}_\gamma = \boldsymbol{U}_\gamma \boldsymbol{D}_\gamma \boldsymbol{V}_\gamma$. We could just use the latent parameter $\boldsymbol{\delta}_\gamma = \boldsymbol{V}_\gamma \boldsymbol{\beta}_\gamma$ of reduced dimension $k_\gamma = n-1$ instead of $\boldsymbol{\beta}_\gamma = \boldsymbol{V}_\gamma^T \boldsymbol{\delta}_\gamma$. Defining the corresponding design matrix as $\boldsymbol{Z}_\gamma = \boldsymbol{W}^{-1/2}\boldsymbol{U}_\gamma \boldsymbol{D}_\gamma$, we have $\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma = \boldsymbol{Z}_\gamma \boldsymbol{\delta}_\gamma$ and retain $\boldsymbol{Z}_\gamma^T \mathbf{1}_n = \mathbf{0}_{k_\gamma}$. Assigning

the prior distribution $\boldsymbol{\delta}_\gamma \sim \mathrm{N}_{k_\gamma}(\mathbf{0}_{k_\gamma}, g\phi c \boldsymbol{D}_\gamma^{-2})$ then induces a normal prior on $\boldsymbol{\beta}_\gamma$ with mean zero and singular precision $(g\phi c)^{-1}\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma$, and thus directly generalizes (6). Investigation of this approach for GLMs with many covariates is another possibility for future research.

Appendix

Proof of prior mode zero

Consider the density function from (5). Dropping for brevity the notational dependency on the model γ, it can be rewritten as

$$f(\boldsymbol{\beta} \,|\, g, \boldsymbol{y}_0) \propto \exp\left\{ \frac{1}{g\phi} \boldsymbol{w}^T \big(h(0)\boldsymbol{\theta} - b(\boldsymbol{\theta}) \big) \right\}, \tag{29}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ and $b(\boldsymbol{\theta}) = \big(b(\theta_1), \ldots, b(\theta_n)\big)^T$. To prove that the mode is at $\boldsymbol{\beta} = \mathbf{0}_p$, note that this is a solution of the score equation

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\boldsymbol{\beta} \,|\, g, \boldsymbol{y}_0) = \frac{1}{g\phi} \left(h(0) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^T} - \frac{\partial\, b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^T} \right)^T \boldsymbol{w} = \mathbf{0}_p,$$

because $\boldsymbol{\beta} = \mathbf{0}_p$ implies that $b'(\theta_i) \equiv b'(\theta) = \mu = h(0)$ and hence

$$\frac{\partial\, b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \mathrm{diag}\big(b'(\theta_1), \ldots, b'(\theta_n)\big) = h(0)\boldsymbol{I}_n.$$

Higher-order Laplace approximation

Denote the standard Laplace approximation (14) by $\tilde{f}_{LA}(\boldsymbol{y} \,|\, g, \gamma)$. Then Raudenbush et al. (2000, p. 148) show that

$$f(\boldsymbol{y} \,|\, g, \gamma) \approx \tilde{f}_{LA}(\boldsymbol{y} \,|\, g, \gamma) \left[1 - \frac{1}{8} \sum_{i=1}^n d_i^{(3)} b_i^2 - \frac{1}{48} \sum_{i=1}^n d_i^{(6)} b_i^3 + \frac{5}{24} \boldsymbol{k}^T (\boldsymbol{R}_{0\gamma}^*)^{-1} \boldsymbol{k} \right] \tag{30}$$

is a sixth-order Laplace approximation when the canonical response function is used. Here $d_i^{(m)} = d^m h / d\eta^m(\eta_i^*)$ evaluated at $\eta_i^* = \boldsymbol{x}_{0\gamma i}^T \boldsymbol{\beta}_{0\gamma}^*$, $b_i = \boldsymbol{x}_{0\gamma i}^T (\boldsymbol{R}_{0\gamma}^*)^{-1} \boldsymbol{x}_{0\gamma i}$ and $\boldsymbol{k} = \sum_{i=1}^n d_i^{(2)} b_i \boldsymbol{x}_{0\gamma i}$. Note that the quadratic forms can be efficiently computed using the Cholesky decomposition $\boldsymbol{R}_{0\gamma}^* = \boldsymbol{L}\boldsymbol{L}^T$, e. g. $\boldsymbol{k}^T (\boldsymbol{R}_{0\gamma}^*)^{-1} \boldsymbol{k} = \|\boldsymbol{v}\|^2$ where $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{k}$.

References

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *Annals of Statistics*, 32(3): 870–897. 398

Berger, J. O. and Pericchi, L. R. (2001). "Objective Bayesian methods for model selection: introduction and comparison." *Lecture Notes-Monograph Series*, 38(1): 135–207. 388

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons. 389

Box, G. E. P. and Tidwell, P. W. (1962). "Transformation of the independent variables." *Technometrics*, 4(4): 531–550. 399

Breiman, L. and Friedman, J. H. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American Statistical Association*, 80(391): 580–598. 396

Brent, R. P. (1973). *Algorithms for Minimization Without Derivatives*. Prentice-Hall series in automatic computation. Englewood Cliffs, NJ: Prentice-Hall. 394

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). "Geographically weighted regression—modelling spatial non-stationarity." *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(3): 431–443. 390

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). "Model selection: an integral part of inference." *Biometrics*, 53(2): 603–618. 398

Chen, M. and Ibrahim, J. (2003). "Conjugate priors for generalized linear models." *Statistica Sinica*, 13: 461–476. 389, 391

Chen, M.-H., Huang, L., Ibrahim, J. G., and Kim, S. (2008). "Bayesian variable selection and computation for generalized linear models with conjugate priors." *Bayesian Analysis*, 3(3): 585–614. 391

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output." *Journal of the American Statistical Association*, 96(453): 270–281. 395

Clyde, M. and George, E. I. (2004). "Model uncertainty." *Statistical Science*, 19(1): 81–94. 387

Cottet, R., Kohn, R. J., and Nott, D. J. (2008). "Variable selection and model averaging in semiparametric overdispersed generalized linear models." *Journal of the American Statistical Association*, 103(482): 661–671. 402

Cui, W. and George, E. I. (2008). "Empirical Bayes vs. fully Bayes variable selection." *Journal of Statistical Planning and Inference*, 138(4): 888–900. 388, 389, 396

Dobra, A. (2009). "Variable selection and dependency networks for genomewide data." *Biostatistics*, 10(4): 621–639. 404

Fernández, C., Ley, E., and Steel, M. F. J. (2001). "Benchmark priors for Bayesian model averaging." *Journal of Econometrics*, 100(2): 381–427. 388

Frank, A. and Asuncion, A. (2010). *UCI Machine Learning Repository*. URL http://archive.ics.uci.edu/ml 396

Gamerman, D. (1997). "Sampling from the posterior distribution in generalized linear mixed models." *Statistics and Computing*, 7(1): 57–68. 393, 395

George, E. I. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87(4): 731–747. 388

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 398

Golub, G. and Welsch, J. (1969). "Calculation of Gauss quadrature rules." *Mathematics of Computation*, 23(106): 221–230. 394

Gupta, M. and Ibrahim, J. (2009). "An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data." *Statistica Sinica*, 19(4): 1641–1663. 391, 392

Han, C. and Carlin, B. (2001). "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review." *Journal of the American Statistical Association*, 96(455): 1122–1132. 395

Hans, C., Dobra, A., and West, M. (2007). "Shotgun stochastic search for "large p" regression." *Journal of the American Statistical Association*, 102(478): 507–516. 404

Hansen, M. H. and Yu, B. (2001). "Model selection and the principle of minimum description length." *Journal of the American Statistical Association*, 96(454): 746–774. 388

— (2003). "Minimum description length model selection criteria for generalized linear models." *Lecture Notes-Monograph Series*, 40(1): 145–163. Statistics and Science: A Festschrift for Terry Speed. 391, 392

Holmes, C. C. and Held, L. (2006). "Bayesian auxiliary variable models for binary and multinomial regression." *Bayesian Analysis*, 1(1): 145–168. 399

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press, third edition. 401

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. 398

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90(431): 928–934. 391

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. 388, 398, 402

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 44(1–2): 187–192. 388

— (1980). "Approximate Bayesian methods." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, 223–245. Valencia: University of Valencia Press. 393

Madigan, D. and York, J. (1995). "Bayesian graphical models for discrete data." *International Statistical Review*, 63(2): 215–232. 401

Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer texts in Statistics. New York: Springer. 392

Maruyama, Y. and George, E. I. (2010). "gBF: A Fully Bayes Factor with a Generalized g-prior." Technical report, Center for Spatial Information Science, University of Tokyo.
URL http://arxiv.org/abs/0801.4410 404

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman and Hall, second edition. 387

Naylor, J. C. and Smith, A. F. M. (1982). "Applications of a method for the efficient computation of posterior distributions." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3): 214–225. 394

Nott, D. J., Kohn, R. J., and Fielding, M. (2008). "Approximating the marginal likelihood using copula." Technical report, Department of Statistics and Applied Probability, National University of Singapore.
URL http://arxiv.org/abs/0810.5474 395

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). "Bayesian variable and link determination for generalised linear models." *Journal of Statistical Planning and Inference*, 111(1-2): 165–180. 391

Overstall, A. M. and Forster, J. J. (2010). "Default Bayesian model determination methods for generalised linear mixed models." *Computational Statistics and Data Analysis*, 54(12): 3269–3288. 391

Pfeffermann, D. (1993). "The role of sampling weights when modeling survey data." *International Statistical Review*, 61(2): 317–337. 390

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 3rd edition. 394

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 394

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). "Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation." *Journal of Computational and Graphical Statistics*, 9(1): 141–157. 393, 405

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press. 396

Robert, C. P. (2001). *The Bayesian Choice*. Springer Texts in Statistics. New York: Springer, second edition. 388

Robert, C. P., Chopin, N., and Rousseau, J. (2009). "Harold Jeffreys's Theory of Probability revisited." *Statistical Science*, 24(2): 141–172. 388

Robert, C. P. and Saleh, A. K. M. E. (1991). "Point estimation and confidence set estimation in a parallelism model: an empirical Bayes approach." *Annales d'Économie et de Statistique*, 23: 65–89. 388

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). "Marginal structural models and causal inference in epidemiology." *Epidemiology*, 11(5): 550–560. 390

Royston, P. and Altman, D. G. (1994). "Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3): 429–467. 399

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(2): 319–392. 393

Sabanés Bové, D. and Held, L. (2010). "Bayesian fractional polynomials." *Statistics and Computing*. Epub ahead of print, DOI: 10.1007/s11222-010-9170-7. 396, 401

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *Annals of Statistics*, 38(5): 2587–2619. 398

Smyth, G., Hu, Y., and Dunn, P. (2010). *statmod: Statistical Modeling*. R package version 1.4.8. 394

Tierney, L. and Kadane, J. B. (1986). "Accurate approximations for posterior moments and marginal densities." *Journal of the American Statistical Association*, 81(393): 82–86. 393

Wang, X. and George, E. I. (2007). "Adaptive Bayesian criteria in variable selection for generalized linear models." *Statistica Sinica*, 17(2): 667–690. 392

Wedderburn, R. W. M. (1976). "On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models." *Biometrika*, 63(1): 27–32. 390

West, M. (1985). "Generalized linear models: scale parameters, outlier accommodation and prior distributions." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, 531–558. Amsterdam: North-Holland. 393

— (2003). "Bayesian factor regression models in the "large p, small n" paradigm." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, 733–742. Oxford University Press. 404

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In Goel, P. K. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, chapter 5, 233–243. Amsterdam: North-Holland. 388

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, 585–603. Valencia: University of Valencia Press. 388, 396

Zhang, Z., Jordan, M. I., and Yeung, D. Y. (2009). "Posterior consistency of the Silverman g-prior in Bayesian model choice." In Koller, D., Bengio, Y., Schuurmans, D., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 21. 404

Acknowledgments

The authors would like to thank the referee and the Associate Editor for helpful comments.

PAPER II

Objective Bayesian model selection in generalised additive models with penalised splines

Daniel Sabanés Bové, Leonhard Held & Göran Kauermann

Paper conditionally accepted and revised for *Journal of Computational and Graphical Statistics*.

Objective Bayesian Model Selection in Generalised Additive Models with Penalised Splines

Daniel Sabanés Bové* Leonhard Held* Göran Kauermann†

Abstract

We propose an objective Bayesian approach to the selection of covariates and their penalised splines transformations in generalised additive models. The methodology is based on a combination of continuous mixtures of g-priors for model parameters and a multiplicity-correction prior for the models themselves. We introduce our approach in the normal model and extend it to non-normal exponential families. A simulation study and an application with binary outcome is provided. An efficient implementation is available in the `R`-package "`hypergsplines`".

Keywords: variable selection, function selection, g-prior, shrinkage, stochastic search

1 Introduction

Semiparametric regression has achieved an impressive dissemination over the last years. Its central idea is to replace parametric regression functions by smooth, semiparametric components. Following Hastie and Tibshirani (1990), suppose we have p continuous

*Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland. E-mail: {daniel.sabanesbove,leonhard.held}@ifspm.uzh.ch

†Department of Statistics, Ludwig-Maximilians-Universität München, Germany. E-mail: goeran.kauermann@stat.uni-muenchen.de

covariates $x_1, \ldots, x_p$ and use the additive model

$$y = \beta_0 + \sum_{j=1}^{p} m_j(x_j) + \epsilon, \tag{1}$$

where the $m_j(\cdot)$, $1 \leq j \leq p$, are smooth but otherwise unspecified functions and $\epsilon \sim \mathrm{N}(0, \sigma^2)$. For identifiability purposes we further assume that $\mathbb{E}\{m_j(X_j)\} = 0$ with respect to the marginal distribution of each covariate X_j. Estimation of the smooth terms in (1) can be carried out in different ways, where we here make use of penalised splines, see *e.g.* Eilers and Marx (2010) or Wood (2006). A general introduction to penalised spline smoothing has been provided by Ruppert, Wand, and Carroll (2003) and the approach has become a popular smoothing technique since then. The general idea is to decompose the functions m_j into a linear and a nonlinear part, where the latter is represented through a spline basis, that is

$$m_j(x_j) = x_j\beta_j + \boldsymbol{Z}_j(x_j)^T \boldsymbol{u}_j. \tag{2}$$

Here $\boldsymbol{Z}_j(x_j)$ is a $K \times 1$ spline basis vector at position x_j and $\boldsymbol{u}_j$ is the corresponding coefficient vector, for $1 \leq j \leq p$. Conveniently one may choose a truncated polynomial basis for $\boldsymbol{Z}_j(\cdot)$ but representation (2) holds in general as well, see Wand and Ormerod (2008). To achieve a smooth fit one imposes a quadratic penalty on the spline coefficient vector $\boldsymbol{u}_j$. Equivalently, one may formulate the penalty as a normal prior

$$\boldsymbol{u}_j \,|\, \sigma^2, \rho_j \sim \mathrm{N}_K(\boldsymbol{0}_K, \sigma^2\rho_j \boldsymbol{I}_K), \tag{3}$$

where $\boldsymbol{0}_K$ is the all-zeros vector and $\boldsymbol{I}_K$ is the identity matrix of dimension K, which leads together with (1) and (2) to a linear mixed model (see Wand, 2003; Kauermann, Krivobokova, and Fahrmeir, 2009). The variance factor ρ_j plays the role of a smoothing parameter which steers the amount of penalisation (relative to the regression variance σ^2). A larger ρ_j leads to a higher prior variance of the spline coefficients and hence a more wiggly function m_j, while a smaller ρ_j leads to a stronger penalty on $\|\boldsymbol{u}_j\|$ and thus a smoother function m_j. In the extreme case, setting ρ_j to zero imposes $\boldsymbol{u}_j \equiv \boldsymbol{0}_K$ so that $m_j(x_j)$ collapses to a linear term $m_j(x_j) = x_j\beta_j$. Hence the role of ρ_j ($j = 1, \ldots, p$) can be seen twofold. For $\rho_j > 0$ it plays the role of a smoothing parameter but with

$\rho_j = 0$ it extends to model selection of (generalised) additive models by separating linear from non-linear effects. We will extend the idea in this paper coherently by proposing a general model selection including variable selection, that is by allowing the alternative $m_j(x_j) \equiv 0$. The central idea is that ρ_j determines uniquely the contribution of the function $m_j(x_j)$ to the overall degrees of freedom of the model (see Ruppert et al., 2003), which is a measure of the complexity of the model. So instead of estimating or drawing inference about ρ_j we draw inference about the corresponding degrees of freedom.

The selection of variables and covariates, respectively, is a central question in statistics. This applies in particular to regression models where the intention is to reduce the variance of effect estimates due to uninformative covariates. The field is wide and many different approaches have been proposed in the last years including the following. Friedman (2001) and Tutz and Binder (2006) describe boosting algorithms, which are extended by Kneib, Hothorn, and Tutz (2009) to geoadditive regression models (Fahrmeir, Kneib, and Lang, 2004). For the same model class, Belitz and Lang (2008) propose to use information-criteria or cross-validation, while Fahrmeir, Kneib, and Konrath (2010) and Scheipl, Fahrmeir, and Kneib (2012) use spike-and-slab priors for variable and function selection (see also Scheipl, Kneib, and Fahrmeir (2013) for simulation studies comparing their approach to the one presented in this paper). Brezger and Lang (2008) adopt the concept of Bayesian contour probabilities (Held, 2004) to decide on the inclusion and form of covariate effects. Cottet, Kohn, and Nott (2008) generalise earlier work by Yau, Kohn, and Wood (2003) to Bayesian double-exponential regression models, which comprise generalised additive models as a special case. Shrinkage approaches are proposed by Wood (2011) and Marra and Wood (2011). Zhang and Lin (2006) use a lasso-type penalised likelihood approach, and Ravikumar, Liu, Lafferty, and Wasserman (2008) and Meier, van de Geer, and Bühlmann (2009) use penalties favouring both sparsity and smoothness of high-dimensional models. Likelihood-ratio testing methods are described by Kauermann and Tutz (2001) and Cantoni and Hastie (2002). This list mirrors the multitude as well as the variety of the different approaches and is, of course, in no way exhaustive.

In this paper we propose a novel objective Bayesian variable and function selection

approach based on continuous mixtures of (generalised) g-priors. This type of prior for the parameters in the generalised additive model traces back to the g-prior in the linear model (Zellner, 1986). Its hyper-parameter g acts as an inverse relative prior sample size, and assigning it a hyper-prior solves the information paradox (Liang, Paulo, Molina, Clyde, and Berger, 2008, section 4.1) of the fixed-g prior (Berger and Pericchi, 2001, p. 148) in the linear model. One specific example are the hyper-g priors of Liang et al. (2008, section 3.2), which enjoy a closed form for the marginal likelihood and lead to consistent model selection and model-averaged prediction. We will proceed to use hyper-g priors, because they have been well studied and have shown good frequentist properties in the Gaussian linear model. They have recently been extended to generalised linear models by Sabanés Bové and Held (2011b). We follow the conventional prior approach (Berger and Pericchi, 2001, section 2.1) by using non-informative improper priors for parameters which are common to all models, and default proper hyper-g priors for model-specific parameters.

While hyper-g priors have been discussed extensively in the Bayesian variable selection literature, *e.g.* by Cui and George (2008), Liang et al. (2008), Bayarri, Berger, Forte, and García-Donato (2012) and Celeux, Anbari, Marin, and Robert (2012), this is the first paper to our knowledge that applies hyper-g priors to generalised additive models. The general idea of applying hyper-g priors, originally developed for linear models, to generalised additive models is new. The rationale is that default priors have carefully and exhaustively been constructed for the linear model, so their advantages should be used when drawing inferences about generalised additive models. Moreover, we consider both variable selection and transformation in a coherent Bayesian framework.

The paper is organised as follows. We first describe how to approach additive models in Section 2, including the specification of hyper-g priors in this model class (Section 2.1), and a suitable multiplicity-correction prior as well as a stochastic search procedure on the model space (Section 2.2). We illustrate the performance of the methodology with a simulation study (Section 2.3). We then extend our focus to generalised additive models in Section 3, which is complemented by an application to real data (Section 3.2). Section 4 closes the paper with a discussion.

2 Additive Models

Assume we have observed independent responses y_i at covariate values $x_{i1}, \ldots, x_{ip}$, $i = 1, \ldots, n$, from the additive normal model (1). For each covariate $j = 1, \ldots, p$, we stack the covariate values into the $n \times 1$ vector $\tilde{\boldsymbol{x}}_j = (x_{1j}, \ldots, x_{nj})^T$ and the spline basis vectors into the $n \times K$ matrix $\tilde{\boldsymbol{Z}}_j = (\boldsymbol{Z}_j(x_{1j}), \ldots, \boldsymbol{Z}_j(x_{nj}))^T$. To achieve orthogonality we apply the Gram-Schmidt process (see Björck, 1967)

$$\boldsymbol{x}_j = \tilde{\boldsymbol{x}}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \tilde{\boldsymbol{x}}_j}{\mathbf{1}_n^T \mathbf{1}_n} = \tilde{\boldsymbol{x}}_j - \mathbf{1}_n \bar{x}_j, \tag{4}$$

$$\boldsymbol{Z}_j = \tilde{\boldsymbol{Z}}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \tilde{\boldsymbol{Z}}_j}{\mathbf{1}_n^T \mathbf{1}_n} - \boldsymbol{x}_j \frac{\boldsymbol{x}_j^T \tilde{\boldsymbol{Z}}_j}{\boldsymbol{x}_j^T \boldsymbol{x}_j}, \tag{5}$$

where $\mathbf{1}_n$ denotes the all-ones vector of dimension n. This ensures that $\mathbf{1}_n$, $\boldsymbol{x}_j$ and the columns of $\boldsymbol{Z}_j$ are orthogonal to each other, *i. e.* $\mathbf{1}_n^T \boldsymbol{x}_j = 0$ and $\mathbf{1}_n^T \boldsymbol{Z}_j = \boldsymbol{x}_j^T \boldsymbol{Z}_j = \mathbf{0}_K$. The orthogonalisation procedure ensures that we can separate the linear and nonlinear part of m_j, which is a prerequisite for the definition of the degrees of freedom measure below. Note that covariates may still be mutually correlated.

A common measure of model complexity is the degrees of freedom of a model. While in parametric models this is just the number of parameters, for smoothing and mixed models Aerts, Claeskens, and Wand (2002, section 2.2) relate the smoothing parameter ρ_j to the corresponding degrees of freedom through

$$d_j(\rho_j) = \operatorname{tr}\{(\boldsymbol{Z}_j^T \boldsymbol{Z}_j + \rho_j^{-1} \boldsymbol{I})^{-1} \boldsymbol{Z}_j^T \boldsymbol{Z}_j\} + 1 \in (1, K+1) \tag{6}$$

for a smoothly modelled covariate effect m_j. Note that $d_j(\rho_j) = \sum_{k=1}^K \lambda_{jk} / (\lambda_{jk} + \rho_j^{-1})$ is easy to calculate via the (positive) eigenvalues λ_{jk} of $\boldsymbol{Z}_j^T \boldsymbol{Z}_j$. This also shows that $d_j(\rho_j)$ is strictly increasing in ρ_j with derivative $\sum_{k=1}^K \lambda_{jk} / (\rho_j \lambda_{jk} + 1)^2 > 0$. This in turn implies that we may (numerically) invert the function to $\rho_j(d_j)$, which means that we have a one-to-one relation between ρ_j and the degrees of freedom d_j. Note that (6) is an asymptotic approximation of the more commonly used definition of degrees of freedom for linear smoothers (see Aerts et al., 2002) and may thus lead to an imprecise measure of model complexity in small samples.

Subsequently we will restrict the degrees of freedom to take values in a finite set $\mathcal{D} \subset \{0\} \cup [1, K+1)$. In the remainder of this article we will use $\mathcal{D} = \{0, 1, 2, 3, \ldots, K\}$, which determines the size of $\mathcal{D}$ to be $K+1$. In general you may want to pick the grid of degrees of freedom to be finer or with the maximum degrees of freedom less than K (perhaps to be chosen by the user), which might be advantageous in some cases. For $d_j = 0$ we set $m_j(x_j) \equiv 0$ while for $d_j = 1$ we have the linear model $m_j(x_j) = x_j\beta_j$. In general, we translate the structure of model (1) into the index vector $\boldsymbol{d} = (d_1, \ldots, d_p)$ giving the degrees of freedom for each functional component. The objective of the paper is to draw inference about $\boldsymbol{d}$, which we subsequently refer to as the "model". To do so, we look now at the stochastic model for the response based on a specific model $\boldsymbol{d}$.

After combining the $I = \sum_{j=1}^{p} \mathbb{I}(d_j \geq 1)$ vectors $\boldsymbol{x}_j$ to the $n \times I$ linear design matrix $\boldsymbol{X}_{\boldsymbol{d}} = (\boldsymbol{x}_j : d_j \geq 1)$ and the $J = \sum_{j=1}^{p} \mathbb{I}(d_j > 1)$ matrices $\boldsymbol{Z}_j$ to the $n \times JK$ spline design matrix $\boldsymbol{Z}_{\boldsymbol{d}} = (\boldsymbol{Z}_j : d_j > 1)$, and analogously constructing the respective coefficient vectors $\boldsymbol{\beta}_{\boldsymbol{d}}$ and $\boldsymbol{u}_{\boldsymbol{d}}$, the conditional additive model for the response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \boldsymbol{u}_{\boldsymbol{d}}, \sigma^2 \sim \mathrm{N}_n \left(\mathbf{1}_n \beta_0 + \boldsymbol{X}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}} + \boldsymbol{Z}_{\boldsymbol{d}} \boldsymbol{u}_{\boldsymbol{d}},\, \sigma^2 \boldsymbol{I}_n \right). \tag{7}$$

Integrating out the the spline coefficient vector $\boldsymbol{u}_{\boldsymbol{d}} \,|\, \sigma^2, \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_{JK}(\mathbf{0}_{JK}, \sigma^2 \boldsymbol{D}_{\boldsymbol{d}})$, where $\boldsymbol{\rho}_{\boldsymbol{d}} = (\rho_j : d_j > 1)$ and $\boldsymbol{D}_{\boldsymbol{d}}$ is block-diagonal with J blocks $\rho_j \boldsymbol{I}_K$ $(d_j > 1)$, yields the so-called marginal model

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \sigma^2, \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_n \left(\mathbf{1}_n \beta_0 + \boldsymbol{X}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}},\, \sigma^2 \boldsymbol{V}_{\boldsymbol{d}} \right) \tag{8}$$

with $\boldsymbol{V}_{\boldsymbol{d}} = \boldsymbol{I}_n + \boldsymbol{Z}_{\boldsymbol{d}} \boldsymbol{D}_{\boldsymbol{d}} \boldsymbol{Z}_{\boldsymbol{d}}^T$. To illustrate the notation, consider for example $p = 4$ covariates and $K = 3$ knots, and a model with degrees of freedom $d_1 = 0$, $d_2 = 1$ and $d_3, d_4 = 2$. Then $\boldsymbol{X}_{\boldsymbol{d}} = (\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$ has $I = 3$ columns, $\boldsymbol{Z}_{\boldsymbol{d}} = (\boldsymbol{Z}_3, \boldsymbol{Z}_4)$ has is composed of $J = 2$ matrices and has $JK = 6$ columns, and $\boldsymbol{D}_{\boldsymbol{d}} = \mathrm{diag}(\rho_3 \boldsymbol{I}_3, \rho_4 \boldsymbol{I}_3)$. This general linear model can be decorrelated into a standard linear model by using the Cholesky decomposition $\boldsymbol{V}_{\boldsymbol{d}} = \boldsymbol{V}_{\boldsymbol{d}}^{T/2} \boldsymbol{V}_{\boldsymbol{d}}^{1/2}$: For the transformed response vector $\tilde{\boldsymbol{y}} = \boldsymbol{V}_{\boldsymbol{d}}^{-T/2} \boldsymbol{y}$ we have

$$\tilde{\boldsymbol{y}} \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \sigma^2, \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_n \left(\tilde{\mathbf{1}}_n \beta_0 + \tilde{\boldsymbol{X}}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}},\, \sigma^2 \boldsymbol{I}_n \right) \tag{9}$$

with analogously transformed all-ones vector $\tilde{\mathbf{1}}_n = \boldsymbol{V}_d^{-T/2}\mathbf{1}_n$ and design matrix $\tilde{\boldsymbol{X}}_d = \boldsymbol{V}_d^{-T/2}\boldsymbol{X}_d$. Note that now also $\tilde{\boldsymbol{y}}$ and $\tilde{\mathbf{1}}_n$ depend on the model $\boldsymbol{d}$, but we suppress this dependence for ease of notation.

2.1 Hyper-g Priors for Additive Models

We will now impose priors on the parameters and show how to use hyper-g priors for the parameter components β_0, $\boldsymbol{\beta}_d$ and σ^2 in the decorrelated marginal model (9). The hyper-g priors comprise a locally uniform prior $f(\beta_0) \propto 1$ on the intercept, Jeffreys' prior $f(\sigma^2) \propto (\sigma^2)^{-1}$ on the regression variance and the g-prior (Zellner, 1986)

$$\boldsymbol{\beta}_d \,|\, g, \sigma^2, \boldsymbol{\rho}_d \sim \mathrm{N}_I\left(\mathbf{0}_I,\, g\sigma^2(\tilde{\boldsymbol{X}}_d^T\tilde{\boldsymbol{X}}_d)^{-1}\right) \tag{10}$$

on the linear coefficient vector. Note that the prior precision matrix in (10) is proportional to $\sigma^{-2}\tilde{\boldsymbol{X}}_d^T\tilde{\boldsymbol{X}}_d = \sigma^{-2}\boldsymbol{X}_d^T\boldsymbol{V}_d^{-1}\boldsymbol{X}_d$, which is the Fisher information matrix of $\boldsymbol{\beta}_d$ in model (8). The prior construction is completed with either a uniform hyper-prior on the shrinkage coefficient $g/(1+g)$,

$$\frac{g}{1+g} \sim \mathrm{U}(0,1), \tag{11}$$

leading to the hyper-g prior, or with

$$\frac{g/n}{1+g/n} \sim \mathrm{U}(0,1), \tag{12}$$

leading to the hyper-g/n prior (Liang et al., 2008). We recommend to use the latter, because it also leads to consistent posterior model probabilities if the true model is the null model (Liang et al., 2008, theorem 4), see Table 1 in Section 2.3 for illustration.

Basically all formulae given by Liang et al. (2008) carry over to our setting, since inner products of the response vector $\boldsymbol{y}$, the all-ones vector $\mathbf{1}_n$ and the design matrix $\boldsymbol{X}_d$ in model (8) carry over to their transformed counterparts $\tilde{\boldsymbol{y}}$, $\tilde{\mathbf{1}}_n$ and $\tilde{\boldsymbol{X}}_d$ in model (9). This is due to

$$\boldsymbol{V}_d^{-1} = (\boldsymbol{I}_n + \boldsymbol{Z}_d\boldsymbol{D}_d\boldsymbol{Z}_d^T)^{-1} = \boldsymbol{I}_n - \boldsymbol{Z}_d(\boldsymbol{Z}_d^T\boldsymbol{Z}_d + \boldsymbol{D}_d^{-1})^{-1}\boldsymbol{Z}_d^T, \tag{13}$$

which follows from the matrix inversion lemma (see Henderson and Searle, 1981) and leads to $\tilde{\mathbf{1}}_n^T\tilde{\mathbf{1}}_n = \mathbf{1}_n^T\mathbf{1}_n = n$, $\tilde{\mathbf{1}}_n^T\tilde{\boldsymbol{X}}_{\boldsymbol{d}} = \mathbf{1}_n^T\boldsymbol{X}_{\boldsymbol{d}} = \mathbf{0}_I$ and $\tilde{\mathbf{1}}_n^T\tilde{\boldsymbol{y}} = \mathbf{1}_n^T\boldsymbol{y}$ by straightforward calculations. A most convenient property of the hyper-g priors is that they yield closed form marginal likelihoods, which need to be computed on the original response scale via the change of variables formula:

$$f(\boldsymbol{y} \mid \boldsymbol{d}) \propto f(\tilde{\boldsymbol{y}} \mid \boldsymbol{d})|\boldsymbol{V}_{\boldsymbol{d}}^{1/2}|^{-1}, \tag{14}$$

where $f(\tilde{\boldsymbol{y}} \mid \boldsymbol{d})$ is the marginal likelihood of the transformed response vector $\tilde{\boldsymbol{y}}$ in the standard linear model (9). The closed forms for $f(\tilde{\boldsymbol{y}} \mid \boldsymbol{d})$ under the hyper-g priors are given in Appendix A.

For completeness we note that other hyper-priors could be assigned to g as well, but they will typically not lead to a closed form of the marginal likelihood. Examples are the incomplete inverse-gamma prior on $1 + g$ (Cui and George, 2008, p. 891), which generalises the above uniform prior on $g/(1 + g)$, and an inverse-gamma prior on g, which corresponds to the Cauchy prior of Zellner and Siow (1980). The hyper-g/n prior is a special case of the robust prior proposed by Bayarri et al. (2012), for which a closed form of the marginal likelihood exists. An overview of hyper-g priors is given by Ley and Steel (2012).

Posterior inference in a given model $\boldsymbol{d}$ is based on Monte Carlo estimation of the parameters in model (7). We therefore use the factorisation

$$f(\beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \boldsymbol{u}_{\boldsymbol{d}}, \sigma^2, g \mid \boldsymbol{y}) = f(\boldsymbol{u}_{\boldsymbol{d}} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \sigma^2, \boldsymbol{y}) f(\beta_0, \boldsymbol{\beta}_{\boldsymbol{d}} \mid \sigma^2, g, \boldsymbol{y}) f(\sigma^2 \mid \boldsymbol{y}) f(g \mid \boldsymbol{y}). \tag{15}$$

Sampling of g, σ^2 and subsequently $\beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}$ can be done along the lines of Sabanés Bové and Held (2011a, section 2.3): Based on the decorrelated model (9), we sample g using inverse sampling (either with a closed-form quantile function, if the hyper-g prior (11) is used, or with a numerical approximation of the quantile function, if the hyper-g/n prior (12) is used), σ^2 from an inverse-gamma distribution, and finally $\beta_0, \boldsymbol{\beta}_{\boldsymbol{d}} \mid g, \sigma^2$ from

a Gaussian distribution. Finally, the spline coefficient vector $\boldsymbol{u}_d$ is sampled from

$$\begin{aligned} f(\boldsymbol{u}_d \mid \beta_0, \boldsymbol{\beta}_d, \sigma^2, \boldsymbol{y}) &\propto f(\boldsymbol{u}_d \mid \sigma^2) f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d, \sigma^2) \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2} \left[\boldsymbol{u}_d^T \boldsymbol{D}_d^{-1} \boldsymbol{u}_d + \|\boldsymbol{y} - \mathbf{1}_n \beta_0 - \boldsymbol{X}_d \boldsymbol{\beta}_d - \boldsymbol{Z}_d \boldsymbol{u}_d\|^2 \right] \right\} \\ &\propto \mathrm{N}_{JK}\left(\boldsymbol{u}_d \mid \boldsymbol{\Sigma}_d \boldsymbol{Z}_d^T (\boldsymbol{y} - \boldsymbol{X}_d \boldsymbol{\beta}_d),\, \sigma^2 \boldsymbol{\Sigma}_d \right), \end{aligned} \tag{16}$$

where $\boldsymbol{\Sigma}_d = (\boldsymbol{Z}_d^T \boldsymbol{Z}_d + \boldsymbol{D}_d^{-1})^{-1}$ and β_0 disappears because $\boldsymbol{Z}_d^T \mathbf{1}_n = \mathbf{0}_{JK}$. A more detailed description of the parameter sampling approach can be found in the supplementary material.

The general intention though is to draw inference about $\boldsymbol{d}$, which is with the prerequisites introduced so far possible as proposed in the next section.

2.2 Model Prior and Stochastic Search

First we propose a prior $f(\boldsymbol{d})$ on the model space $\mathcal{D}^p$ which explicitly corrects for the multiplicity of testing inherent in the simultaneous analysis of the p covariates (see Scott and Berger, 2010): *A priori*, the number of covariates included in the model (I) is uniformly distributed on $\{0, 1, \ldots, p\}$. The choice of the I covariates is then uniformly distributed on all possible configurations, and their degrees of freedom are independent and uniformly distributed on $\mathcal{D} \setminus \{0\} = \{1, 2, 3, \ldots, K\}$. Altogether, this gives

$$1/f(\boldsymbol{d}) = (p+1) \binom{p}{I} K^I. \tag{17}$$

A nice property of this prior is that it leads to marginal prior probabilities $\mathbb{P}(d_j = 0) = \mathbb{P}(d_j > 0) = 1/2$. Elsewhere this is often achieved by assigning independent priors to the p covariates, which implies that averaged over all models, $I \sim \mathrm{Bin}(p, 1/2)$. It is clear that our uniform prior on I allows the data $\boldsymbol{y}$ to have a maximum effect on the posterior of I because it is the reference prior (Bernardo, 1979). Note that this prior actually favours models with high or low numbers of covariates, as there are fewer such models. This or similar model priors have been used in a number of previous papers, including *e. g.* George and McCulloch (1993) and Ley and Steel (2009).

Alternatively, one might also use a fixed (independent of K) prior probability for a linear effect ($d_j = 1$). This is appropriate for the situation where one explicitly wants to

test linearity versus nonlinearity of each effect. Furthermore, a multiplicity correction for these tests can be implemented by assuming that the number of smoothly included covariates (J) is uniformly distributed on $\{0, 1, \ldots, I\}$ and their choice is uniform on all possible choices. This would add one level to the prior hierarchy.

As the model space $\mathcal{D}^p$ grows exponentially in the number of covariates p, only for small values of p all possible models can be evaluated. Otherwise the marginal likelihoods $f(\boldsymbol{y} \mid \boldsymbol{d})$ and posterior model probabilities $f(\boldsymbol{d} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{d}) f(\boldsymbol{d})$ can be computed only for a subset of the model space. Usually this subset is determined by stochastic search procedures (Madigan and York, 1995). Here we propose to use a simple Metropolis-Hastings algorithm with two possible move types in the proposal kernel:

Move Sample a covariate index $j \sim \mathrm{U}\{1, 2, \ldots, p\}$ and decrease or increase d_j to the next adjacent value in $\mathcal{D}$ (with probability $1/2$ each, or deterministically if $d_j = 0$ or $d_j = K$, respectively).

Swap Sample a pair $(i, j) \sim \mathrm{U}\{(1,1), (1,2), \ldots, (p,p)\}$ of covariate indices ($i \leq j$) and swap d_i and d_j.

The 'Swap' move is designed to efficiently trace models with high posterior probability even in situations where covariates are almost collinear. For each Metropolis-Hastings iteration, a 'Move' is chosen with some fixed probability (we use $3/4$), and otherwise a 'Swap'. Denote the current model by $\boldsymbol{d}$, then the proposed model $\boldsymbol{d}'$ is accepted with probability

$$\alpha(\boldsymbol{d}' \mid \boldsymbol{d}) = 1 \wedge \frac{f(\boldsymbol{y} \mid \boldsymbol{d}') f(\boldsymbol{d}') q(\boldsymbol{d}' \mid \boldsymbol{d})}{f(\boldsymbol{y} \mid \boldsymbol{d}) f(\boldsymbol{d}) q(\boldsymbol{d} \mid \boldsymbol{d}')}$$

where the calculation of the proposal probability ratio $q(\boldsymbol{d}' \mid \boldsymbol{d}) / q(\boldsymbol{d} \mid \boldsymbol{d}')$ is straightforward (see the supplementary material).

2.3 Simulation Study

In order to study the performance of our approach in identifying the true model, we performed a simulation study. Full details are provided in the supplementary material; Here we summarise the main results. Three different true models were simulated: The

first model ("null") was the null model with $p = 20$ nuisance covariates. The second model ("small") also had $p = 20$ covariates of which 3 had a linear effect and 3 had a nonlinear (quadratic, sine, and skew-normal density) effect. Correlations of different strength were generated between some of the covariates. The third model ("large") was identical to the second model, but included additional 80 nuisance covariates, which were independent of the first 20 covariates. For the "small" and "large" models, one covariate was chosen to be a surrogate for the true (quadratic) effect of another covariate. It masks the quadratic effect if only linear effects can be fitted by a variable selection algorithm. For three different sample sizes $n = 50, 100, 1000$, and for the three different true models, we simulated n observations from the Gaussian additive model (1) with $\beta_0 = 0$ and $\sigma^2 = 0.2^2$. This was repeated 50 times for each combination of model and sample size, in order to assess the sampling variability.

We applied the proposed additive model selection approaches to each data set, using the hyper-priors (11) and (12) ("hyper-g splines" and "hyper-g/n splines", respectively). As the computational complexity of the marginal likelihood (14) is cubic in the spline basis dimension K (see the supplementary material), we want to use splines with few, quantile-based knots. Therefore, we choose cubic O'Sullivan splines (Wand and Ormerod, 2008). Here, we got basis matrices $\boldsymbol{Z}_j$ with $K = 8$ columns from 6 inner knots at the septiles. We applied the stochastic search algorithm described in Section 2.2 with 10^6 iterations.

We compared the results with those from pure variable selection including only linear functions ("hyper-g linear" and "hyper-g/n linear"), Bayesian fractional polynomials ("Bayesian FPs") (Sabanés Bové and Held, 2011a), spike-and-slab function selection ("Spike-and-slab", Scheipl et al., 2012) and splines knot selection ("Knot selection", Denison, Mallick, and Smith, 1998, using code from chapters 3 and 4 in Denison, Holmes, Mallick, and Smith, 2002).

Concerning the discovery of the true set of influential covariates, the additive model selection procedures introduced in this paper were very competitive with the considered alternative methods, as is illustrated in Table 1. In particular, they showed clear advantages in the case of small and moderate sample sizes. Using splines instead of

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	$83_{(8)}$	$84_{(7)}$	$84_{(9)}$	$49_{(25)}$	$65_{(13)}$	$86_{(14)}$	$2_{(26)}$	$74_{(15)}$	$87_{(16)}$
Hyper-g/n splines	$86_{(10)}$	$91_{(6)}$	$97_{(3)}$	$47_{(24)}$	$68_{(14)}$	$87_{(13)}$	$0_{(24)}$	$75_{(15)}$	$89_{(15)}$
Hyper-g linear	$20_{(7)}$	$21_{(6)}$	$23_{(7)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$
Hyper-g/n linear	$50_{(16)}$	$64_{(15)}$	$90_{(8)}$	$0_{(0)}$	$0_{(1)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$	$0_{(0)}$
Bayesian FPs	$37_{(6)}$	$37_{(7)}$	$37_{(6)}$	$2_{(15)}$	$35_{(16)}$	$3_{(23)}$	$0_{(9)}$	$47_{(19)}$	$37_{(28)}$
Spike-and-slab	$89_{(6)}$	$93_{(2)}$	$98_{(0)}$	$3_{(5)}$	$45_{(6)}$	$79_{(2)}$	$0_{(0)}$	$10_{(8)}$	$71_{(5)}$
Knot selection	$92_{(8)}$	$94_{(5)}$	$98_{(2)}$	$0_{(1)}$	$34_{(20)}$	$95_{(6)}$	$0_{(0)}$	$0_{(1)}$	$89_{(9)}$

Table 1 *– Median posterior probability of the true model in percentage, when the true model is defined by correct variable inclusion. Standard deviations (in parentheses) are computed from the 50 replications.*

only linear functions proved essential for the discovery of the masked quadratic effect and hence convergence to the true model. Looking at the standard deviations in the 50 replications, we observe for the hyper-g and hyper-g/n spline methods a relatively high variability for $n = 50$, which decreases then for larger sample sizes. Interestingly the variability is increasing for the Bayesian FPs, and no clear trend is visible for the spike-and-slab and knot selection methods.

Variable inclusion performance did not differ substantively with respect to sensitivity, specificity and area under the ROC curve between the considered methods, with the exception of a slightly worse performance of the two linear methods. However, as shown in Table 2, the hyper-g and hyper-g/n spline methods were clearly better in distinguishing the truly effective covariates from the highly correlated nuisance covariates. Moreover, for small sample sizes, they outperformed the other nonlinear methodologies concerning the discovery of the masked quadratic effect. In this task the merely linear methods obviously failed. With respect to sampling variability, the proposed spline methods are very competitive, with smallest variability among all methods for larger sample sizes.

Concerning the average mean squared errors of the model-averaged posterior mean

	small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	$75_{(29)}$	$97_{(3)}$	$98_{(3)}$	$26_{(36)}$	$100_{(0)}$	$100_{(0)}$
Hyper-g/n splines	$79_{(26)}$	$97_{(3)}$	$98_{(3)}$	$20_{(50)}$	$100_{(0)}$	$100_{(0)}$
Hyper-g linear	$18_{(17)}$	$44_{(19)}$	$87_{(8)}$	$6_{(11)}$	$26_{(33)}$	$98_{(3)}$
Hyper-g/n linear	$22_{(21)}$	$48_{(22)}$	$90_{(8)}$	$17_{(44)}$	$26_{(33)}$	$98_{(2)}$
Bayesian FPs	$41_{(33)}$	$89_{(12)}$	$68_{(16)}$	$9_{(18)}$	$92_{(19)}$	$81_{(15)}$
Spike-and-slab	$30_{(19)}$	$88_{(5)}$	$97_{(0)}$	$1_{(2)}$	$60_{(19)}$	$97_{(1)}$
Knot selection	$9_{(19)}$	$78_{(22)}$	$99_{(1)}$	$4_{(11)}$	$13_{(20)}$	$99_{(3)}$

Table 2 – *Average difference* $\frac{1}{2}(P_{16} + P_{17}) - \frac{1}{3}(P_{18} + P_{19} + P_{20})$ *of inclusion probabilities* $P_j = \mathbb{P}\{m_j(x_j) \neq 0 \,|\, \boldsymbol{y}\}$ *(in percentage points) between the truly effective covariates* x_{16} *and* x_{17} *and the nuisance covariates* x_{18}, x_{19}, x_{20}*, which had correlation* 0.8 *with* x_{16} *and* x_{17}*. (The optimal value is* 100*, the worst value is* -100*.) Standard deviations (in parentheses) are computed from the 50 replications.*

function estimates $\hat{m}_j(x_j)$, the proposed additive model selection procedures were very competitive. They performed well or better than the best compared methods each, as is shown in Table 3. It is interesting that the hyper-g splines were slightly but consistently better than the hyper-g/n splines. We also investigated the coverage rates of pointwise 95% credible intervals for the functions, and found that the two proposed methods were slightly conservative.

Finally, the average computational effort of the two proposed additive model selection procedures ranged between one minute for $n = 100$ in a "null" data set to about 50 minutes for $n = 50$ in a "large" data set (times to be expected on a 2.8 GHz single-core CPU, see the supplementary material for more details).

Average MSE	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	344.38	114.14	22.43	39.15	10.32	1.68	30.42	1.88	0.33
Hyper-g/n splines	462.92	71.72	2.17	47.82	18.33	3.20	784.44	2.78	0.61
Hyper-g linear	7586.25	1378.49	137.06	158.10	133.55	121.97	45.11	32.26	24.36
Hyper-g/n linear	2155.39	182.62	6.78	189.57	169.00	120.96	378.07	36.23	26.09
Bayesian FPs	1424.17	283.76	19.20	16837.92	3026.61	29.51	76.78	356.30	5.80
Spike-and-slab	19038.78	18224.91	5660.40	80.94	14.00	2.09	45.45	8.71	0.81
Knot selection	337.77	36.79	0.65	180.03	35.29	2.07	47.23	29.33	0.78

Standard deviation	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	1268.26	341.24	114.91	36.32	3.49	0.26	17.89	0.51	0.05
Hyper-g/n splines	1435.00	169.16	8.59	28.22	6.05	1.20	1720.04	0.88	0.19
Hyper-g linear	20871.36	2197.63	208.82	23.28	20.47	5.28	8.15	5.19	1.10
Hyper-g/n linear	6172.70	446.47	33.01	27.02	28.45	4.55	589.21	4.61	1.59
Bayesian FPs	5760.39	973.94	38.44	118315.78	21154.38	5.63	248.39	2471.83	0.97
Spike-and-slab	8768.03	5619.08	1762.70	30.05	6.43	0.40	5.13	3.60	0.09
Knot selection	1734.24	126.34	2.35	39.28	28.25	0.40	6.89	4.59	0.32

Table 3 – *Average mean squared errors (top table, in 10^{-8} units for the "null" model, and 10^{-4} units for the "small" and "large" models) and corresponding standard deviations (bottom table, same units as in top table) of function estimates. Numbers are averaged over all covariates and the 50 replications, standard deviations are computed from the 50 replications.*

3 Generalised Additive Models

Now we extend the above setting and assume that the covariate effects $m_j(x_j)$ enter additively into the linear predictor

$$\eta = \beta_0 + \sum_{j=1}^{p} m_j(x_j) \tag{18}$$

of an exponential family distribution with canonical parameter θ, mean $\mathbb{E}(y) = h(\eta) = db(\theta)/d\theta$ and variance $\mathrm{Var}(y) = \phi/w \cdot d^2b(\theta)/d\theta^2$ (see McCullagh and Nelder, 1989). We restrict our attention to non-normal distributions with fixed dispersion ϕ (as $\phi = 1$ for the Bernoulli and Poisson distribution) and known weight w. For n observations, the linear predictor vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ is

$$\boldsymbol{\eta} = \mathbf{1}_n\beta_0 + \boldsymbol{X}_{\boldsymbol{d}}\boldsymbol{\beta}_{\boldsymbol{d}} + \boldsymbol{Z}_{\boldsymbol{d}}\boldsymbol{u}_{\boldsymbol{d}} \tag{19}$$

and the likelihood is

$$f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \boldsymbol{u}_{\boldsymbol{d}}) \propto \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/w_i}\right\}. \tag{20}$$

The main challenge for the derivation of a generalised g-prior is that the marginal density $f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}})$, which results from integrating out the spline coefficient vector

$$\boldsymbol{u}_{\boldsymbol{d}} \mid \boldsymbol{\rho}_{\boldsymbol{d}} \sim \mathrm{N}_{JK}(\mathbf{0}_{JK}, \boldsymbol{D}_{\boldsymbol{d}}) \tag{21}$$

from (20), has no closed form. In particular, it is not Gaussian, in contrast to (8).

Calculation of the degrees of freedom $d_j(\rho_j)$ for a smoothly modelled term m_j can be carried out with a reasonable generalisation of (6), that is (see Ruppert et al., 2003, section 11.4)

$$d_j(\rho_j) = \mathrm{tr}\{(\boldsymbol{Z}_j^T\widehat{\boldsymbol{W}}\boldsymbol{Z}_j + \rho_j^{-1}\boldsymbol{I})^{-1}\boldsymbol{Z}_j^T\widehat{\boldsymbol{W}}\boldsymbol{Z}_j\} + 1, \tag{22}$$

which uses a fixed weight matrix $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\mathbf{1}_n\widehat{\beta}_0)$, where $\boldsymbol{W}(\boldsymbol{\eta}) = \mathrm{diag}\{(dh(\eta_i)/d\eta)^2/\,\mathrm{Var}(y_i)\}_{i=1}^n$ is the usual generalised linear model weight matrix and $\widehat{\beta}_0$ is the intercept estimate from the null model $\boldsymbol{d} = \mathbf{0}_p$. This definition avoids dependence of $\rho_j(d_j)$ on the model $\boldsymbol{d}$ under consideration and serves as simplification. In particular it again allows to invert $d_j(\rho_j)$

to obtain the variance component ρ_j for a given degree d_j. As a consequence, we next need to generalise the orthogonalisation of the original covariate vector $\tilde{\boldsymbol{x}}_j$ and spline basis matrix $\tilde{\boldsymbol{Z}}_j$ from (4) and (5) to

$$\boldsymbol{x}_j = \tilde{\boldsymbol{x}}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \tilde{\boldsymbol{x}}_j}{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \mathbf{1}_n} \tag{23}$$

$$\text{and} \quad \boldsymbol{Z}_j = \tilde{\boldsymbol{Z}}_j - \mathbf{1}_n \frac{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \tilde{\boldsymbol{Z}}_j}{\mathbf{1}_n^T \widehat{\boldsymbol{W}} \mathbf{1}_n} - \boldsymbol{x}_j \frac{\boldsymbol{x}_j^T \widehat{\boldsymbol{W}} \tilde{\boldsymbol{Z}}_j}{\boldsymbol{x}_j^T \widehat{\boldsymbol{W}} \boldsymbol{x}_j}, \tag{24}$$

implying that $\mathbf{1}_n$, $\boldsymbol{x}_j$ and the columns of $\boldsymbol{Z}_j$ are orthogonal to each other with respect to the inner product in terms of $\widehat{\boldsymbol{W}}$. This ensures again that (22) correctly captures only the degrees of freedom associated with the nonlinear part of m_j.

3.1 Hyper-g Priors for Generalised Additive Models

We will now derive a generalised g-prior analogous to (10) for the linear coefficient vector $\boldsymbol{\beta}_{\boldsymbol{d}}$ in the generalised additive model. The idea is to apply the iterative weighted least squares (IWLS) approximation to the non-normal likelihood (20) to obtain an approximate normal model of the form (7) and then derive the resulting g-prior (10). With a slight abuse of notation, *e. g.* $h(\boldsymbol{\eta}) = (h(\eta_1), \ldots, h(\eta_n))^T$, let

$$\boldsymbol{z}_0 = \boldsymbol{\eta}_0 + \operatorname{diag}\{dh(\boldsymbol{\eta}_0)/d\boldsymbol{\eta}\}^{-1}\{\boldsymbol{y} - h(\boldsymbol{\eta}_0)\} \tag{25}$$

be the adjusted response vector resulting from a first-order approximation to $h^{-1}(\boldsymbol{y})$ around $\boldsymbol{y} = h(\boldsymbol{\eta}_0)$. Then

$$\boldsymbol{z}_0 \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \boldsymbol{u}_{\boldsymbol{d}} \overset{\text{approx}}{\sim} \mathrm{N}\big(\mathbf{1}_n \beta_0 + \boldsymbol{X}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}} + \boldsymbol{Z}_{\boldsymbol{d}} \boldsymbol{u}_{\boldsymbol{d}},\, \boldsymbol{W}_0^{-1}\big) \tag{26}$$

with $\boldsymbol{W}_0 = \boldsymbol{W}(\boldsymbol{\eta}_0)$ is the working normal model (see *e. g.* McCullagh and Nelder, 1989, p. 40). The IWLS algorithm iteratively updates $\boldsymbol{\eta}_0$ by weighted least squares estimation of the coefficients in (26). Here, we fix $\boldsymbol{\eta}_0 = \mathbf{0}_n$, which is the value expected *a priori*. Then we rewrite (26) using $\tilde{\boldsymbol{z}}_0 = \boldsymbol{W}_0^{1/2} \boldsymbol{z}_0$, $\tilde{\mathbf{1}}_n = \boldsymbol{W}_0^{1/2} \mathbf{1}_n$, $\tilde{\boldsymbol{X}}_{\boldsymbol{d}} = \boldsymbol{W}_0^{1/2} \boldsymbol{X}_{\boldsymbol{d}}$ and $\tilde{\boldsymbol{Z}}_{\boldsymbol{d}} = \boldsymbol{W}_0^{1/2} \boldsymbol{Z}_{\boldsymbol{d}}$ as

$$\tilde{\boldsymbol{z}}_0 \,|\, \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \boldsymbol{u}_{\boldsymbol{d}} \overset{\text{approx}}{\sim} \mathrm{N}(\tilde{\mathbf{1}}_n \beta_0 + \tilde{\boldsymbol{X}}_{\boldsymbol{d}} \boldsymbol{\beta}_{\boldsymbol{d}} + \tilde{\boldsymbol{Z}}_{\boldsymbol{d}} \boldsymbol{u}_{\boldsymbol{d}},\, \boldsymbol{I}_n), \tag{27}$$

which brings us back to a normal model of the form in (7). By computing the corresponding g-prior (10), we arrive at the generalised g-prior

$$\boldsymbol{\beta}_d \,|\, g, \boldsymbol{\rho}_d \sim \mathrm{N}_I(\mathbf{0}_I, g\boldsymbol{J}_0^{-1}) \tag{28}$$

with prior precision matrix proportional to

$$\begin{aligned} \boldsymbol{J}_0 &= \tilde{\boldsymbol{X}}_d^T(\boldsymbol{I}_n + \tilde{\boldsymbol{Z}}_d\boldsymbol{D}_d\tilde{\boldsymbol{Z}}_d^T)^{-1}\tilde{\boldsymbol{X}}_d \\ &= \boldsymbol{X}_d^T\boldsymbol{W}_0^{1/2}(\boldsymbol{I}_n + \boldsymbol{W}_0^{1/2}\boldsymbol{Z}_d\boldsymbol{D}_d\boldsymbol{Z}_d^T\boldsymbol{W}_0^{1/2})^{-1}\boldsymbol{W}_0^{1/2}\boldsymbol{X}_d. \end{aligned} \tag{29}$$

An appealing feature of this prior is that it directly generalises the g-prior proposed by Sabanés Bové and Held (2011b) for generalised linear models, to which it reduces when there are no spline effects in the model, *i.e.* $\boldsymbol{J}_0 = \boldsymbol{X}_d^T\boldsymbol{W}_0\boldsymbol{X}_d$. An alternative and more rigorous derivation of (29) as the Fisher information obtained from a Laplace approximation to the marginal model $f(\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_d)$ is provided in Appendix B.

The generalised hyper-g prior

$$f(\beta_0, \boldsymbol{\beta}_d, \boldsymbol{u}_d, g) = f(\beta_0)f(\boldsymbol{\beta}_d \,|\, g, \boldsymbol{\rho}_d)f(g)f(\boldsymbol{u}_d) \tag{30}$$

is defined to comprise the locally uniform prior $f(\beta_0) \propto 1$ on the intercept β_0, the generalised g-prior (28) on the linear coefficient vector $\boldsymbol{\beta}_d$, the penalty prior (21) on the spline coefficient vector $\boldsymbol{u}_d$, and some proper hyper-prior $f(g)$ on the hyper-parameter g. Posterior inference under this prior can be implemented as outlined in the following. The efficient `R`-package "`hypergsplines`" for this and all other computations in this paper is available from `R`-Forge at `http://hypergsplines.r-forge.r-project.org/`. For installation, just type `install.packages("hypergsplines",repos="http://r-forge.r-project.org")` into `R`.

Let $\boldsymbol{X}_a = (\mathbf{1}_n, \boldsymbol{X}_d, \boldsymbol{Z}_d)$ and $\boldsymbol{\beta}_a = (\beta_0, \boldsymbol{\beta}_d^T, \boldsymbol{u}_d^T)^T$ denote the grand design matrix and regression coefficient vector, respectively, such that $\boldsymbol{\eta} = \boldsymbol{X}_a\boldsymbol{\beta}_a$. The prior for $\boldsymbol{\beta}_a$ conditional on g has a Gaussian form with mean zero and singular precision matrix $\mathrm{diag}(0, g^{-1}\boldsymbol{J}_0, \boldsymbol{D}_d^{-1})$. Thus, the Gaussian approximation of $f(\boldsymbol{\beta}_a \,|\, \boldsymbol{y}, g, \boldsymbol{d})$, which is necessary for the Laplace approximation of $f(\boldsymbol{y} \,|\, g, \boldsymbol{d})$, can be obtained by the Bayesian IWLS algorithm (West,

1985). Afterwards, an approximation of the marginal likelihood of model $\boldsymbol{d}$,

$$f(\boldsymbol{y} \mid \boldsymbol{d}) = \int_0^\infty f(\boldsymbol{y} \mid g, \boldsymbol{d}) f(g)\, dg, \tag{31}$$

is obtained by numerical integration of the Laplace approximation $\tilde{f}(\boldsymbol{y} \mid g, \boldsymbol{d})$. For small sample sizes, using a higher order Laplace approximation can be useful, see Sabanés Bové and Held (2011b, section 3.1). Note that integrated Laplace approximations have successfully been applied in a more general context (Rue, Martino, and Chopin, 2009). Finally, we can use a tuning-free Metropolis-Hastings algorithm to sample from the joint posterior of $\boldsymbol{\beta}_a$ and g in a specific model $\boldsymbol{d}$: Values g are sampled on the log-scale from a proposal density obtained by linear interpolation of pairs $\{z_j, \tilde{f}(z_j, \boldsymbol{y} \mid \boldsymbol{d})\}$, $j = 1, \ldots, 20$, which are already used for the above numerical integration of the Laplace approximation. Here $\tilde{f}(z, \boldsymbol{y} \mid \boldsymbol{d}) = \tilde{f}(\boldsymbol{y} \mid g, \boldsymbol{d}) f(g) g$ is the approximated unnormalised posterior density of $z = \log(g)$. Note that this sampling scheme for g can be interpreted as an approximate griddy Gibbs sampler (Ritter and Tanner, 1992). Conditional on the proposed value of g, a Gaussian proposal density for $\boldsymbol{\beta}_a$ is obtained by performing one or more IWLS steps from the previous state of $\boldsymbol{\beta}_a$ (Gamerman, 1997). See Sabanés Bové and Held (2011b, section 3), on which this implementation is based on, for more details on the computations.

3.2 Application

We now apply the generalised additive model selection approach to the logistic regression of $p = 7$ potential risk factors on the presence of diabetes in $n = 532$ women of Pima Indian heritage (Ripley, 1996; Frank and Asuncion, 2010), see Table 4 for details. We use cubic O'Sullivan splines with 4 inner knots at the quintiles and the hyper-prior (12), and explore the model space of dimension $7^7 = 823\,543$ with 10^6 iterations of the stochastic search algorithm. Note that the most complex model spends $4 \cdot 7 = 28$ degrees of freedom. Considering the recommendation that a parametric logistic regression model should contain at least 10 events (successes or failures) for each independent explanatory variable (Peduzzi, Concato, Kemper, Holford, and Feinstein, 1996), this most complex

model would be large because we only have 177 successes in this data set. This rule easily extends to nonparametric logistic regression by replacing the number of explanatory variables by the total degrees of freedom. From this perspective it is not recommended to use more knots for the spline bases. More knots also do not change the results in this example, as we have seen when using 9 inner knots at the deciles.

The computational complexity is higher than for the normal response case, with 95 minutes required for the evaluation of the 39 081 models found. We validated the results with an exhaustive evaluation of all models, requiring 33 hours. Indeed, the stochastic search found 99% of the posterior probability mass and the 733 top models.

Variable	Description
y	Signs of diabetes according to WHO criteria (Yes = 1, No = 0)
x_1	Number of pregnancies
x_2	Plasma glucose concentration in an oral glucose tolerance test [mg/dl]
x_3	Diastolic blood pressure [mm Hg]
x_4	Triceps skin fold thickness [mm]
x_5	Body mass index (BMI) [kg/m^2]
x_6	Diabetes pedigree function
x_7	Age [years]

Table 4 – *Description of the variables in the Pima Indian diabetes data set. Note that the original dataset has $n = 768$ observations and $p = 8$ explanatory variables, but several missing values. We dropped the variable* `insulin` *with the highest proportion of missing values and removed the remaining rows with missing data to perform a complete case analysis.*

In Table 5 the marginal posterior probabilities for linear and smooth inclusion of the covariates are shown. There is clear evidence for inclusion of the covariates x_2, x_5, x_6 and x_7, which have posterior inclusion probabilities over 96%. For the other three covariates, the inclusion probability is below 30%. Smooth modelling of the effects of x_5, x_6 and x_7 seems to be necessary, while this is not so clear for x_2.

In order to examine the mixing properties of the stochastic search algorithm proposed

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
not included ($d_j = 0$)	0.74	0.00	0.88	0.91	0.00	0.04	0.01
linear ($d_j = 1$)	0.07	0.48	0.06	0.04	0.11	0.26	0.00
smooth ($d_j > 1$)	0.19	0.52	0.06	0.05	0.89	0.70	0.99

Table 5 *– Marginal posterior inclusion probabilities in the Pima Indian diabetes data set.*

in Section 2.2, we compared the results based on starting the MCMC chain from the full model with $d_j = 4$ instead of the previously used null model with $d_j = 0$ ($j = 1, \ldots, p$). The results are very close: for example, the entries in Table 5 differ by at most $2.28 \cdot 10^{-4}$, and the top 500 models which were visited by the chains are identical. These results are an indication that slow mixing is not a problem for the presented stochastic search algorithm for this example. It is recommended to perform similar checks for all applications.

Figure 1 shows the estimated covariate effects in the *maximum a posteriori* (MAP) model which features a linear term for x_2 and smooth terms for x_5, x_6 and x_7. The estimates are obtained from 10 000 MCMC samples (every 2nd sample after burning the first 1000 iterations of the Markov chain). Using two IWLS steps per proposal yielded an acceptance rate of 67%. Note that for linear functions m_j, the pointwise credible intervals coincide with the simultaneous credible intervals (Besag, Green, Higdon, and Mengersen, 1995, p. 30). This is because all straight lines samples intersect in one point, which is due to the centring of the covariates in (23). Furthermore, we observe that the Chib and Jeliazkov (2001) estimate (-240.924, MCMC standard error 0.008) of the log marginal likelihood of the MAP model, which was also computed, is quite close to the integrated Laplace approximation (-241.01). This indicates that the integrated Laplace approximation is fairly accurate.

When the main interest lies in variable selection, multiple models which feature the same covariates can be summarised into a single meta-model as follows: The posterior probabilities of the sub-models are summed up to give the posterior probability of the

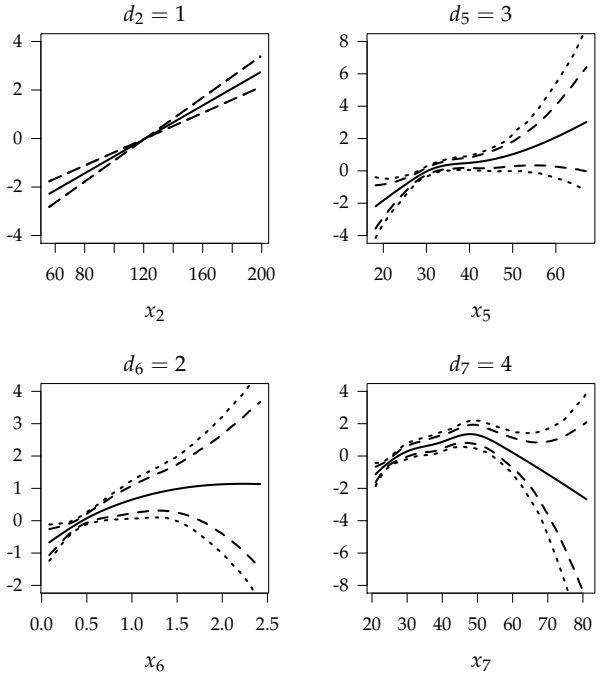


Figure 1 – *Estimated covariate effects in the MAP model for the Pima Indian diabetes data set, based on 10 000 MCMC samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.*

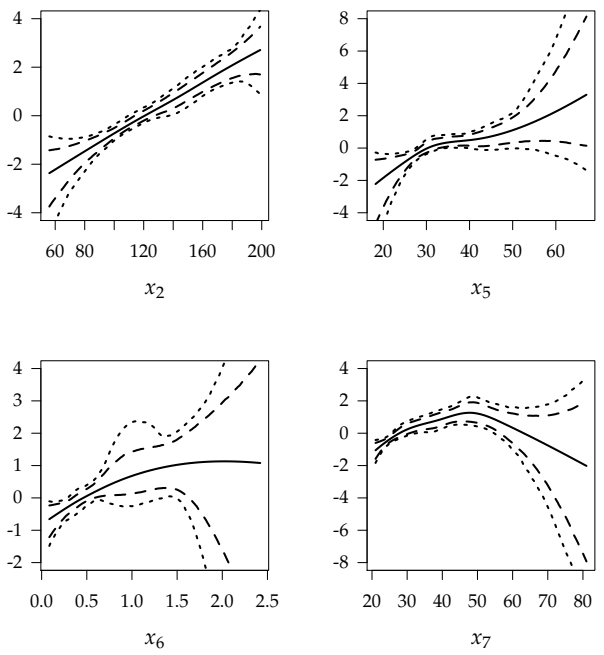


Figure 2 – *Estimated covariate effects in the best meta-model (and median probability meta-model) for the Pima Indian diabetes data, based on 20 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.*

meta-model, and estimates in the meta-model are obtained by averaging the sub-models with weights proportional to their posterior probabilities (see *e. g.* Hoeting, Madigan, Raftery, and Volinsky, 1999, for model averaging). Here the best meta-model includes x_2, x_5, x_6 and x_7 and has posterior probability 0.598. The corresponding estimates of the covariate effects are shown in Figure 2. This best meta-model happens to be identical with the median probability meta-model, which features all covariates having marginal posterior inclusion probability greater than 50% (Barbieri and Berger, 2004), *cp*. Table 5. Similarly, it could be interesting to summarise models which only differ in the degrees of freedom for smooth terms. This would correspond to the situation of testing linearity versus nonlinearity of covariate effects (*cp*. Section 2.2).

In summary, the results are qualitatively similar to those obtained with a FP modelling approach by Sabanés Bové and Held (2011b, section 5) and with a cubic smoothing spline approach by Cottet et al. (2008, section 3.2). It is interesting that in the earlier work

by Yau et al. (2003, section 5.2), a very low posterior inclusion probability (0.07) for x_6 was reported for a different subset of the original Pima Indian diabetes data set. If pure variable selection without covariate transformation is considered, as in Holmes and Held (2006, section 2.6) and Sabanés Bové and Held (2011b, section 4), the strong nonlinear effect of x_7 is missed completely, and instead x_1 gets a higher posterior inclusion probability. This may be a case of a masked nonlinear effect, as was simulated in Section 2.3, and highlights the importance of allowing for nonlinear covariate effects.

4 Discussion

Our Bayesian approach to simultaneous variable and function selection in generalised regression is based on fixed-dimensional spline bases and penalty-parameter smoothness control. In this way it is coherent and differs from knot-selection approaches such as Smith and Kohn (1996) and Denison et al. (1998). We found that fixed-dimensional spline bases based on a small number of knots are flexible enough to capture the functional forms we expect (see *e. g.* Abrahamowicz, MacKenzie, and Esdaile, 1996). Furthermore, at least in the example from Section 3.2, the results are very robust to increasing the number of knots. In the interest of computation times we thus recommend to use only a small number of knots. Moreover, by using fixed-dimensional smooth components we can constrain a covariate effect to be exactly linear. This enables us to look at posterior probabilities of linear *versus* smooth inclusion of covariates. Approaches which use variable-dimensional smooth components and select knots, as Denison et al. (1998), cannot fit linear functions.

We are only considering roughness penalties on a fixed grid of values, which scales automatically for each covariate via the degrees of freedom transformation. We found that it is a very useful approximation of a continuous scale. One possibility for checking the quality of the discrete approximation is to optimise the marginal likelihood of the MAP model with respect to the degrees of freedom of the covariates included. That is, an optimisation of $f(\boldsymbol{y} \mid \boldsymbol{d})$ over the continuous range $1 < d_j < K + 1$ is performed for all covariates included in the MAP model. For example, the MAP configuration for the

Pima Indian diabetes data is $(0, 1, 0, 0, 3, 2, 4)$ and the resulting optimised configuration is $(0, 1, 0, 0, 3.42, 2.1, 3.74)$, which increases the log marginal likelihood from -241.01 to -240.86. Although d_5 and d_7 changed considerably in the optimisation, the resulting function estimates are very similar to those from the MAP model in Figure 1. In all examples we have looked at, the resulting optimised models yielded very similar results compared to the MAP model, which indicates that the fixed grid approximation is good enough. In this regard, our approach is close to many popular Lasso-type proposals, which optimise the tuning-parameters on a fixed grid via cross-validation (*e. g.* Zou and Hastie, 2005). Cantoni and Hastie (2002) propose a likelihood-ratio-type test statistic to compare additive models with different degrees of freedom. Fong, Rue, and Wakefield (2010) use a similar scaling to examine the prior on the degrees of freedom implied by the prior on the variance component in a generalised linear mixed model. They also use O'Sullivan spline bases as we did in our applications, but they do not consider variable selection.

In a frequentist setting, Marra and Wood (2011, section 2.1) propose to use an additional penalty on the linear part of the spline function in order to shrink it adaptively to zero. To include variable selection, a lower threshold for the effective degrees of freedom must be chosen. Our generalised g-prior (28) also shrinks the linear parts of the spline functions to zero, where the prior covariance matrix takes the correlations between the covariates into account. Incorporating the covariates correlation in the coefficients prior allows for better discrimination between influential and correlated nuisance covariates. Empirical results from our simulation study in Section 2.3 support this. Furthermore, we explicitly ex- or include covariates and then compare the resulting models based on their posterior probabilities. This avoids *ad-hoc* choices of a threshold and leads to a coherent variable selection procedure.

We propose a conventional prior for the intercept and the linear coefficients, which directly generalises the hyper-g priors in the linear model (Liang et al., 2008) and in the generalised linear model (Sabanés Bové and Held, 2011b). Pauler (1998) proposes a related unit-information prior for the fixed effects in linear mixed models, but fixes $g = n$ in (10). Overstall and Forster (2010) propose a unit-information prior for the fixed

effects in generalised linear mixed models, but the information matrix is based on the first-stage likelihood and not on the integrated likelihood as in our approach. Also, no hyper-prior on the parameter g is considered, because it is fixed at $g = n$. As they use an inverse-Wishart prior on the covariance matrix of the random effects, their approach is perhaps better suited to generic random effects models. Forster, Gill, and Overstall (2012) propose a novel reversible-jump MCMC algorithm to infer the corresponding posterior model probabilities. We are confident that our proposed generalised additive model selection procedure, which can be used with any of the various well-explored default priors in the linear model, is a competitive alternative to other approaches.

Appendix

Appendix A gives details on the closed form of the marginal likelihood (14) for normal additive models. In Appendix B, an alternative derivation of the prior precision matrix (29) in the generalised g-prior is presented.

A Closed Forms of Marginal Likelihood in Additive Models

Under the hyper-g prior (11), the marginal likelihood of the transformed response vector is (Liang et al., 2008)

$$f(\tilde{\boldsymbol{y}} \mid \boldsymbol{d}) \propto \|\boldsymbol{V}_{\boldsymbol{d}}^{-T/2}(\boldsymbol{y} - \mathbf{1}_n \bar{y})\|^{-(n-1)} (I+2)^{-1} {}_2\mathrm{F}_1\left(\frac{n-1}{2}; 1; \frac{I+4}{2}; \tilde{R}^2_{\boldsymbol{d}}\right) \tag{32}$$

where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$, ${}_2\mathrm{F}_1$ is the Gaussian hypergeometric function (Abramowitz and Stegun, 1964, p. 558) and $\tilde{R}^2_{\boldsymbol{d}}$ is the classical coefficient of determination in model (8). Under the hyper-g/n prior (12), the marginal likelihood in the standard linear model is

(Forte, 2011, p. 155)

$$f(\tilde{\boldsymbol{y}} \mid \boldsymbol{d}) \propto n^{-I/2}(1-\tilde{R}_{\boldsymbol{d}}^2)^{-(n-1)/2}\frac{2}{I+2} \times \mathrm{AF}_1\left(\frac{I}{2}+1; \frac{I+1-n}{2}; \frac{n-1}{2}; \frac{I}{2}+2; \frac{n-1}{n}, \frac{n-(1-\tilde{R}_{\boldsymbol{d}}^2)^{-1}}{n}\right), \tag{33}$$

where AF_1 is the Appell hypergeometric function of the first kind (Appell, 1925). Colavecchia and Gasaneo (2004) provide Fortran code for computing this special function, which is accessible in R via the package "appell" (Sabanés Bové, 2012). For large sample sizes ($n > 100$) or when the numerical computations of the special functions in (32) or (33) fail, we instead use Laplace approximations as described by Liang et al. (2008, Appendix A). See the supplementary material for details on efficient computation of $\tilde{R}_{\boldsymbol{d}}^2$.

B Approximate Fisher Information in Generalised Additive Models

In this section, we present a formal derivation of formula (29) as the approximate Fisher information obtained from a Laplace approximation to $f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}})$. For ease of notation we restrict the presentation to canonical response functions where $\eta = \theta$ and omit subscripts where they are not necessary for understanding. With $\boldsymbol{\Phi} = \mathrm{diag}\{\phi/w_i\}_{i=1}^n$, we can then rewrite the likelihood (20) as

$$f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left\{\boldsymbol{y}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta} - \mathbf{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta})\right\}. \tag{34}$$

We will now use the Laplace approximation to integrate (34) over $\boldsymbol{u}$ with respect to the prior $\boldsymbol{u} \mid \boldsymbol{\rho} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$.

We first need to maximise the unnormalised log posterior of $\boldsymbol{u}$,

$$\begin{aligned} l(\boldsymbol{u}) &= \log\{f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{u})\} + \log\{f(\boldsymbol{u})\} \\ &= \boldsymbol{y}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta} - \mathbf{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta}) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{D}^{-1}\boldsymbol{u} + \text{const}, \end{aligned} \tag{35}$$

where β_0 and $\boldsymbol{\beta}$ in $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$ are considered to be fixed. The corresponding

score vector is

$$\begin{aligned}\frac{d}{d\boldsymbol{u}}l(\boldsymbol{u}) &= \boldsymbol{Z}^T\boldsymbol{\Phi}^{-1}\boldsymbol{y} - \boldsymbol{Z}^T\operatorname{diag}\{b'(\boldsymbol{\eta})\}\boldsymbol{\Phi}^{-1}\mathbf{1} - \boldsymbol{D}^{-1}\boldsymbol{u} \\ &= \boldsymbol{Z}^T\boldsymbol{\Phi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \boldsymbol{D}^{-1}\boldsymbol{u},\end{aligned}$$

where $\boldsymbol{\mu} = b'(\boldsymbol{\eta})$, and the corresponding Hessian is

$$\begin{aligned}\frac{d}{d\boldsymbol{u}}\frac{d}{d\boldsymbol{u}^T}l(\boldsymbol{u}) &= \frac{d}{d\boldsymbol{u}}\left\{(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Phi}^{-1}\boldsymbol{Z} - \boldsymbol{u}^T\boldsymbol{D}^{-1}\right\} \\ &= -\boldsymbol{Z}^T\boldsymbol{W}(\boldsymbol{\eta})\boldsymbol{Z} - \boldsymbol{D}^{-1}.\end{aligned}$$

Making one Newton-Raphson step from the starting point $\boldsymbol{u} = \mathbf{0}$, we get the approximate mode $\boldsymbol{u}^*$ of $l(\boldsymbol{u})$:

$$\begin{aligned}\boldsymbol{u}^* &= \mathbf{0} - \left(\frac{d}{d\boldsymbol{u}}\frac{d}{d\boldsymbol{u}^T}l(\mathbf{0})\right)^{-1}\frac{d}{d\boldsymbol{u}}l(\mathbf{0}) \\ &= \left(\boldsymbol{Z}^T\boldsymbol{W}(\boldsymbol{\eta}_L)\boldsymbol{Z} + \boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{\Phi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_L),\end{aligned} \tag{36}$$

where $\boldsymbol{\eta}_L = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\mu}_L = b'(\boldsymbol{\eta}_L)$. Note that this corresponds to the result of a second-order Taylor expansion of $l(\boldsymbol{u})$ around $\boldsymbol{u} = \mathbf{0}$. Hence, the Laplace approximation of $f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta})$ is

$$\begin{aligned}\tilde{f}(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}) &\propto \exp(l(\boldsymbol{u}^*))(2\pi)^{JK/2}\left|-\frac{d}{d\boldsymbol{u}}\frac{d}{d\boldsymbol{u}^T}l(\boldsymbol{u}^*)\right|^{-1/2} \\ &= \exp\left(\boldsymbol{y}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}^* - \mathbf{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta}^*) - \frac{1}{2}\boldsymbol{u}^{*T}\boldsymbol{D}^{-1}\boldsymbol{u}^*\right) \\ &\quad \times (2\pi)^{JK/2}\left|\boldsymbol{Z}^T\boldsymbol{W}(\boldsymbol{\eta}^*)\boldsymbol{Z} + \boldsymbol{D}^{-1}\right|^{-1/2},\end{aligned} \tag{37}$$

where JK is the dimension of $\boldsymbol{u}$.

In order to derive the approximate Fisher information of $\boldsymbol{\beta}$ from $\tilde{f}(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta})$, we make two additional simplifying assumptions: First, we assume that $\boldsymbol{W}(\boldsymbol{\eta})$ does not vary much in $\boldsymbol{\beta}$, so that we can ignore the determinant in (37), for example. This is a common simplification, suggested *e. g.* in Breslow and Clayton (1993). Second, we approximate $b(\boldsymbol{\eta}^*)$ by a second-order Taylor expansion of $b(\boldsymbol{\eta})$ around $\boldsymbol{\eta}_L$, yielding

$$\mathbf{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta}^*) \approx \mathbf{1}^T\boldsymbol{\Phi}^{-1}b(\boldsymbol{\eta}_L) + \boldsymbol{\mu}_L^T\boldsymbol{\Phi}^{-1}\boldsymbol{Z}\boldsymbol{u}^* + \frac{1}{2}\boldsymbol{u}^{*T}\boldsymbol{Z}^T\boldsymbol{W}_L\boldsymbol{Z}\boldsymbol{u}^*,$$

where $\boldsymbol{W}_L = \boldsymbol{W}(\boldsymbol{\eta}_L)$. Using these two simplifications and plugging in (36), we arrive at the expression

$$\begin{aligned}\log\{\tilde{f}(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta})\} &= \boldsymbol{y}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}_L - \mathbf{1}^T \boldsymbol{\Phi}^{-1} b(\boldsymbol{\eta}_L) \\ &\quad + (\boldsymbol{y} - \boldsymbol{\mu}_L)^T \boldsymbol{\Phi}^{-1} \boldsymbol{Z} \boldsymbol{u}^* - \frac{1}{2} \boldsymbol{u}^{*T} (\boldsymbol{Z}^T \boldsymbol{W}_L \boldsymbol{Z} + \boldsymbol{D}^{-1}) \boldsymbol{u}^* \\ &= \boldsymbol{y}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\eta}_L - \mathbf{1}^T \boldsymbol{\Phi}^{-1} b(\boldsymbol{\eta}_L) \\ &\quad + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu}_L)^T \boldsymbol{\Phi}^{-1} \boldsymbol{Z} (\boldsymbol{Z}^T \boldsymbol{W}_L \boldsymbol{Z} + \boldsymbol{D}^{-1})^{-1} \boldsymbol{Z}^T \boldsymbol{\Phi}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_L) \end{aligned} \tag{38}$$

for the approximate marginal log-likelihood of β_0 and $\boldsymbol{\beta}$. From (38) we can finally approximate the Fisher information $\boldsymbol{J}(\beta_0, \boldsymbol{\beta}) = -\frac{d}{d\boldsymbol{\beta}} \frac{d}{d\boldsymbol{\beta}^T} \log\{f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta})\}$ as

$$\begin{aligned}\tilde{\boldsymbol{J}}(\beta_0, \boldsymbol{\beta}) &= -\frac{d}{d\boldsymbol{\beta}} \frac{d}{d\boldsymbol{\beta}^T} \log\{\tilde{f}(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta})\} \\ &= \boldsymbol{X}^T \boldsymbol{W}_L^{1/2} \left(\boldsymbol{I} - \boldsymbol{W}_L^{1/2} \boldsymbol{Z} (\boldsymbol{Z}^T \boldsymbol{W}_L \boldsymbol{Z} + \boldsymbol{D}^{-1})^{-1} \boldsymbol{Z}^T \boldsymbol{W}_L^{1/2} \right) \boldsymbol{W}_L^{1/2} \boldsymbol{X} \qquad (39) \\ &= \boldsymbol{X}^T \boldsymbol{W}_L^{1/2} (\boldsymbol{I} + \boldsymbol{W}_L^{1/2} \boldsymbol{Z} \boldsymbol{D} \boldsymbol{Z}^T \boldsymbol{W}_L^{1/2})^{-1} \boldsymbol{W}_L^{1/2} \boldsymbol{X}. \qquad (40)\end{aligned}$$

Evaluating the approximate Fisher information at $\beta_0 = 0, \boldsymbol{\beta} = \mathbf{0}$, such that $\boldsymbol{W}_L = \boldsymbol{W}(\mathbf{0})$, we recognise that $\tilde{\boldsymbol{J}}(0, \mathbf{0})$ from (40) is identical to $\boldsymbol{J}_0$ in formula (29). Note that the representation (39) can be better suited for computation: the second paragraph of Section 2.1 in the supplementary material applies here after replacing $\boldsymbol{Z}_d$ with $\boldsymbol{W}_L^{1/2} \boldsymbol{Z}$.

References

Abrahamowicz, M., MacKenzie, T., and Esdaile, J. M. (1996), "Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis," *Journal of the American Statistical Association*, 91, 1432–1439.

Abramowitz, M. and Stegun, I. A. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover, ninth Dover printing, tenth GPO printing ed.

Aerts, M., Claeskens, G., and Wand, M. P. (2002), "Some theory for penalized spline generalized additive models," *Journal of Statistical Planning and Inference*, 103, 455–470.

Appell, M. P. (1925), "Sur les fonctions hypergéométriques de plusieurs variables, les polynomes d'Hermite et autres fonctions spheriques dans l'hyperespace," *Mémorial des sciences mathématiques*, 3, 1–75.

Barbieri, M. M. and Berger, J. O. (2004), "Optimal predictive model selection," *Annals of Statistics*, 32, 870–897.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012), "Criteria for Bayesian model choice with application to variable selection," *Annals of Statistics*, 40, 1550–1577.

Belitz, C. and Lang, S. (2008), "Simultaneous selection of variables and smoothing parameters in structured additive regression models," *Computational Statistics and Data Analysis*, 53, 61–81.

Berger, J. O. and Pericchi, L. R. (2001), "Objective Bayesian methods for model selection: introduction and comparison," in *Model Selection*, ed. Lahiri, P., Beachwood, OH: Institute of Mathematical Statistics, vol. 38 of *IMS Lecture Notes*, pp. 135–207.

Bernardo, J. M. (1979), "Reference posterior distributions for Bayesian inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 113–147.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian computation and stochastic systems (with discussion)," *Statistical Science*, 10, 3–66.

Björck, Å. (1967), "Solving linear least squares problems by Gram-Schmidt orthogonalization," *BIT Numerical Mathematics*, 7, 1–21.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Brezger, A. and Lang, S. (2008), "Simultaneous probability statements for Bayesian P-splines," *Statistical Modelling*, 8, 141–168.

Cantoni, E. and Hastie, T. (2002), "Degrees-of-freedom tests for smoothing splines," *Biometrika*, 89, 251–263.

Celeux, G., Anbari, M. E., Marin, J.-M., and Robert, C. P. (2012), "Regularization in regression: comparing Bayesian and frequentist methods in a poorly informative situation," *Bayesian Analysis*, 7, 477–502.

Chib, S. and Jeliazkov, I. (2001), "Marginal likelihood from the Metropolis-Hastings output," *Journal of the American Statistical Association*, 96, 270–281.

Colavecchia, F. and Gasaneo, G. (2004), "f1: a code to compute Appell's F1 hypergeometric function," *Computer Physics Communications*, 157, 32–38.

Cottet, R., Kohn, R. J., and Nott, D. J. (2008), "Variable selection and model averaging in semiparametric overdispersed generalized linear models," *Journal of the American Statistical Association*, 103, 661–671.

Cui, W. and George, E. I. (2008), "Empirical Bayes vs. fully Bayes variable selection," *Journal of Statistical Planning and Inference*, 138, 888–900.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Wiley Series in Probability and Statistics, Chichester: Wiley.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 333–350.

Eilers, P. H. C. and Marx, B. D. (2010), "Splines, knots, and penalties," *Wiley Interdisciplinary Reviews Computational Statistics*, 2, 637–653.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010), "Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection," *Statistics and Computing*, 20, 203–219.

Fahrmeir, L., Kneib, T., and Lang, S. (2004), "Penalized structured additive regression for space-time data: A Bayesian perspective," *Statistica Sinica*, 14, 715–745.

Fong, Y., Rue, H., and Wakefield, J. (2010), "Bayesian inference for generalized linear mixed models," *Biostatistics*, 11, 397–412.

Forster, J., Gill, R., and Overstall, A. (2012), "Reversible jump methods for generalised linear models and generalised linear mixed models," *Statistics and Computing*, 22, 107–120.

Forte, A. (2011), "Objective Bayes Criteria for Variable Selection," Ph.D. thesis, Universitat de València, available at https://www.educacion.gob.es/teseo/imprimirFicheroTesis.do?fichero=22234.

Frank, A. and Asuncion, A. (2010), "UCI Machine Learning Repository," available at http://archive.ics.uci.edu/ml.

Friedman, J. H. (2001), "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 29, 1189–1232.

Gamerman, D. (1997), "Sampling from the posterior distribution in generalized linear mixed models," *Statistics and Computing*, 7, 57–68.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall.

Held, L. (2004), "Simultaneous posterior probability statements from Monte Carlo output," *Journal of Computational and Graphical Statistics*, 13, 20–35.

Henderson, H. V. and Searle, S. R. (1981), "On deriving the inverse of a sum of matrices," *SIAM Review*, 23, 53–60.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian model averaging: a tutorial," *Statistical Science*, 14, 382–417.

Holmes, C. C. and Held, L. (2006), "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Analysis*, 1, 145–168.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009), "Some asymptotic results on generalized penalized spline smoothing," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71, 487–503.

Kauermann, G. and Tutz, G. (2001), "Testing generalized linear and semiparametric models against smooth alternatives," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63, 147–166.

Kneib, T., Hothorn, T., and Tutz, G. (2009), "Variable selection and model choice in geoadditive regression models," *Biometrics*, 65, 626–634.

Ley, E. and Steel, M. F. (2009), "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression," *Journal of Applied Econometrics*, 24, 651–674.

— (2012), "Mixtures of g-priors for Bayesian model averaging with economic applications," *Journal of Econometrics*, 171, 251–266.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of *g* priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

Madigan, D. and York, J. (1995), "Bayesian graphical models for discrete data," *International Statistical Review*, 63, 215–232.

Marra, G. and Wood, S. N. (2011), "Practical variable selection for generalized additive models," *Computational Statistics and Data Analysis*, 55, 2372–2387.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, no. 37 in Monographs on Statistics and Applied Probability, New York: Chapman and Hall, 2nd ed.

Meier, L., van de Geer, S., and Bühlmann, P. (2009), "High-dimensional additive modeling," *Annals of Statistics*, 37, 3779–3821.

Overstall, A. M. and Forster, J. J. (2010), "Default Bayesian model determination methods for generalised linear mixed models," *Computational Statistics and Data Analysis*, 54, 3269–3288.

Pauler, D. K. (1998), "The Schwarz criterion and related methods for normal linear models," *Biometrika*, 85, 13–27.

Peduzzi, P., Concato, J., Kemper, E., Holford, T., and Feinstein, A. (1996), "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, 49, 1373–1379.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2008), "SpAM: Sparse additive models," in *Advances in Neural Information Processing Systems 20*, eds. Platt, J., Koller, D., Singer, Y., and Roweis, S., Cambridge, MA: MIT Press, pp. 1201–1208.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Ritter, C. and Tanner, M. A. (1992), "Facilitating the Gibbs sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71, 319–392.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press.

Sabanés Bové, D. (2012), *appell: Compute Appell's F1 hypergeometric function*, R package version 0.0-3, available at http://cran.r-project.org/web/packages/appell/.

Sabanés Bové, D. and Held, L. (2011a), "Bayesian fractional polynomials," *Statistics and Computing*, 21, 309–324.

— (2011b), "Hyper-g priors for generalized linear models," *Bayesian Analysis*, 6, 387–410.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012), "Spike-and-slab priors for function selection in structured additive regression models," *Journal of the American Statistical Association*, 107, 1518–1532.

Scheipl, F., Kneib, T., and Fahrmeir, L. (2013), "Penalized likelihood and Bayesian function selection in regression models," *Advances in Statistical Analysis*, to appear.

Scott, J. G. and Berger, J. O. (2010), "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem," *Annals of Statistics*, 38, 2587–2619.

Smith, M. and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, 75, 317–343.

Tutz, G. and Binder, H. (2006), "Generalized additive modeling with implicit variable selection by likelihood-based boosting," *Biometrics*, 62, 961–971.

Wand, M. P. (2003), "Smoothing and mixed models," *Computational Statistics*, 18, 223–249.

Wand, M. P. and Ormerod, J. T. (2008), "On semiparametric regression with O'Sullivan penalized splines," *Australian & New Zealand Journal of Statistics*, 50, 179–198.

West, M. (1985), "Generalized linear models: scale parameters, outlier accommodation and prior distributions," in *Bayesian Statistics 2*, eds. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., Amsterdam: North-Holland, pp. 531–558.

Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, Boca Raton: Chapman & Hall/ CRC.

— (2011), "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73, 3–36.

Yau, P., Kohn, R. J., and Wood, S. (2003), "Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression," *Journal of Computational and Graphical Statistics*, 12, 23–54.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. Goel, P. K. and Zellner, A., Amsterdam: North-Holland, vol. 6 of *Studies in Bayesian Econometrics and Statistics*, chap. 5, pp. 233–243.

Zellner, A. and Siow, A. (1980), "Posterior odds ratios for selected regression hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, eds. Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., Valencia: University of Valencia Press, pp. 585–603.

Zhang, H. H. and Lin, Y. (2006), "Component selection and smoothing for nonparametric regression in exponential families," *Statistica Sinica*, 16, 1021–1041.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, 301–320.

Objective Bayesian Model Selection in Generalised Additive Models with Penalised Splines – Supplementary Material

Daniel Sabanés Bové* Leonhard Held* Göran Kauermann†

This supplement describes in Section 1 in detail the simulation study for evaluating and comparing the performance of the proposed additive model selection approach. In Section 2 we present an additional example. Finally, implementation details for efficient computation of the marginal likelihood, the parameter sampling in a given additive model and the proposal probabilities in the stochastic search procedure are described in Section 3.

1 Simulation Study

1.1 Setup

The data generating process is described in Section 1.1.1, and the different model selection methods are summarised in Section 1.1.2.

*Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland. E-mail: {daniel.sabanesbove,leonhard.held}@ifspm.uzh.ch

†Department of Statistics, Ludwig-Maximilians-Universität München, Germany. E-mail: goeran.kauermann@stat.uni-muenchen.de

1.1.1 Data generating process

Three different true models were simulated: The first model ("null") was the null model with $p = 20$ nuisance covariates. The second model ("small") also had $p = 20$ covariates of which 3 had a linear effect and 3 had a nonlinear effect. The third model ("large") was identical to the second model, but with a total of $p = 100$ covariates.

The covariates were generated as follows. Consider first the "null" and the "small" model. The covariate vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i,20})^T \in [0,1]^{20}$ were generated from the Gaussian copula with uniform marginal distributions and positive-definite covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{I}_5 & & \boldsymbol{\Sigma}_2^T \\ & \boldsymbol{I}_{10} & \\ \boldsymbol{\Sigma}_2 & & \boldsymbol{\Sigma}_1 \end{pmatrix},$$

where $\boldsymbol{I}_k$ is the identity matrix of dimension k, $\boldsymbol{\Sigma}_1 = (1-\rho)\boldsymbol{I}_5 + \rho \mathbf{1}_5 \mathbf{1}_5^T$ is the (5×5) constant-correlation matrix with correlation $\rho = 0.8$, and $\boldsymbol{\Sigma}_2 = \mathbf{1}_5(0.1, 0.2, 0.3, 0.4, 0.5)$ is the (5×5) matrix specifying increasing correlations between the x_j and $\{x_{16}, \ldots, x_{20}\}$, $j = 1, \ldots, 5$. This means, after generating $\boldsymbol{x}_i \overset{ind}{\sim} \mathrm{N}_{20}(\mathbf{0}, \boldsymbol{\Sigma})$, we transformed each value x_{ij} to the unit interval $[0,1]$ via $x_{ij} \leftarrow \Phi(x_{ij})$, where Φ is the standard normal cdf.

For the "large" model with $p = 100$, 80 uniformly distributed nuisance covariates were added to the data set.

For the "small" and "large" model, we distributed the truly effective covariates evenly across the three different correlation groups of covariates x_j ($j = 1, \ldots, 5$, $j = 6, \ldots, 15$ and $j = 16, \ldots, 20$), with one linear and one nonlinear effect per each correlation group. The true functions $m_j(x_j)$ of the effective covariates x_j are plotted in Figure 1. Variable x_8 from the second correlation group was chosen to be a surrogate for the true effect of x_4. It masks the quadratic effect of x_4 if only linear effects can be fitted by the variable selection algorithm. This was done by setting

$$x_8 = 0.5 m_4(x_4) + \epsilon, \quad \epsilon \overset{iid}{\sim} \mathrm{N}(0, 0.5^2),$$

and afterwards scaling x_8 to the unit interval.

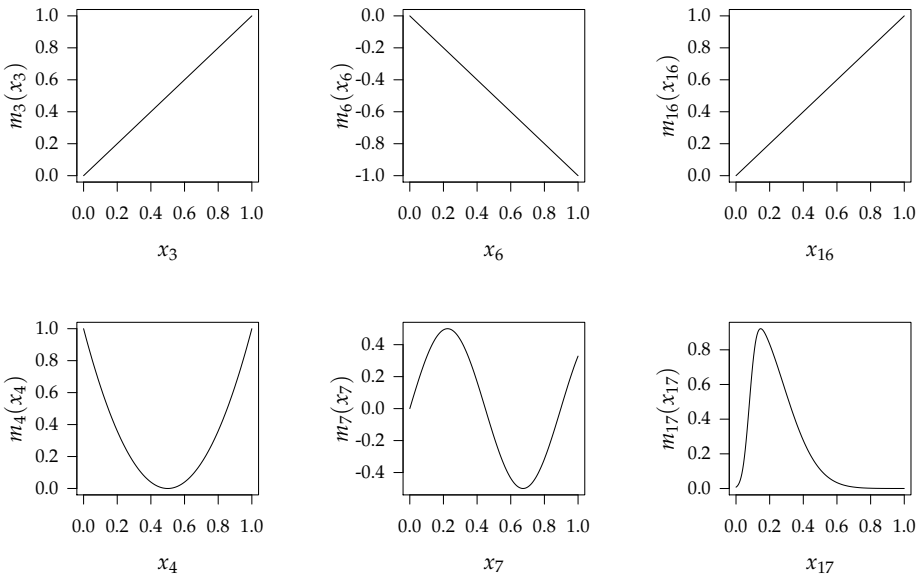


Figure 1 – *True functions of effective covariates in the "small" and "large" model. The linear functions are* $m_3(x) = m_{16}(x) = x$ *and* $m_6(x) = -x$. *The nonlinear functions are* $m_4(x) = 4(x - 1/2)^2$ *(quadratic),* $m_7(x) = 0.5\sin(7x)$ *(sine) and* $m_{17}(x) = 2.5\phi\{(x - 0.08)/0.2\}\Phi\{30(x - 0.08)\}$ *(skew-normal), where* ϕ *and* Φ *are the pdf and cdf of the standard normal distribution, respectively.*

For three different sample sizes $n = 50, 100, 1000$, and for the three different true models, we simulated observations y_i from the Gaussian additive model

$$y_i \overset{ind}{\sim} \mathrm{N}\{\sum_{j=1}^{p} m_j(x_{ij}), 0.2^2\}, \quad i = 1, \ldots, n.$$

This was repeated 50 times for each combination of model and sample size, in order to assess the sampling variability.

1.1.2 Methods

We compared the following methods for variable selection:

1. Hyper-g splines: As described in section 2 of the paper, using the hyper-g prior. We choose cubic O'Sullivan splines (Wand and Ormerod, 2008), and got basis matrices $\boldsymbol{Z}_j$ with $K = 8$ columns from 6 inner knots at the covariate quintiles. We used the stochastic model search with 10^6 iterations.

2. Hyper-g/n splines: Here we used the hyper-g/n prior.

3. Hyper-g linear: We allowed only linear inclusion of covariates as in Liang, Paulo, Molina, Clyde, and Berger (2008), using the hyper-g prior.

4. Hyper-g/n linear: Here we used the hyper-g/n prior.

5. Bayesian fractional polynomials (FPs): As described in Sabanés Bové and Held (2011) and implemented in the `R`-package `bfp`, using the hyper-g prior. We used the stochastic model search with 10^6 iterations and saved the best 3000 models.

6. Spike-and-slab: As described in Scheipl, Fahrmeir, and Kneib (2012) and implemented in the `R`-package `spikeSlabGAM`. We used three parallel chains with a burn-in of 500 samples, saved 2500 MCMC samples and thinned to every second sample. All hyperparameters were set to their default values.

7. Knot selection: As described in Denison, Holmes, Mallick, and Smith (2002, chapters 3 and 4). We translated the corresponding Matlab code written by Chris

Holmes provided at http://www.stat.tamu.edu/~bmallick/wileybook/book_code.html to R code. We used cubic regression splines with a maximum of 100 selected knots in total, saving 3000 MCMC samples after a burn-in of 2000. All other hyperparameters were set to their default values.

We deliberately included methods 3 and 4 which cannot account for nonlinear covariate effects, in order to illustrate the disadvantages in presence of true nonlinear covariate effects.

Methods 1, 2, 5 and 6 can differentiate between linear and nonlinear inclusion of a covariate in the model, while methods 3 and 4 only allow linear, and method 7 only allows nonlinear inclusion of covariates.

1.2 Results

Section 1.2.1 looks at the success in discovering the true model, variable selection performance is analysed in Section 1.2.2, and the quality of function estimates is evaluated in Section 1.2.3. Finally the computational effort of the methods is summarised in Section 1.2.4.

1.2.1 Discovering the true model

Differentiating only correct inclusion of covariates in the model, each of the seven methods assigns a posterior probability to the true model configuration. Taking the median over the 50 replications, we arrive at the numbers in Table 1. (Note that we compute medians here because they are more robust to outliers than means.)

First consider the null model case. We note that using the hyper-g splines/linear or Bayesian FPs, the probabilities do not increase for larger sample size. This is due to the use of the hyper-g prior, which is not consistent under the null model (Liang et al., 2008, section 4.2). For all other methods, the probabilities increase with larger n. It is also interesting that hyper-g/n splines performs better than hyper-g/n linear.

Second consider the small model case. Most striking is the failure of the two linear methods. This is mainly due to the masking of x_4 by x_8. The Bayesian FPs perform

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	83	84	84	49	65	86	2	74	87
Hyper-g/n splines	86	91	97	47	68	87	0	75	89
Hyper-g linear	20	21	23	0	0	0	0	0	0
Hyper-g/n linear	50	64	90	0	0	0	0	0	0
Bayesian FPs	37	37	37	2	35	3	0	47	37
Spike-and-slab	89	93	98	3	45	79	0	10	71
Knot selection	92	94	98	0	34	95	0	0	89

Table 1 – *Median posterior probability of the true model in percentage, when the true model is defined by correct variable inclusion.*

slightly better than the linear methods. The hyper-g and hyper-g/n splines perform similarly and have an advantage over spike-and-slab and knot selection for the smaller sample size $n = 50$.

Finally, in the large model case, all methods perform poorly with only $n = 50$ observations, while for larger sample sizes the methods from this paper perform best.

Instead of looking at the posterior model probability of the true model, one can also look at the posterior rank of the true model. We report the medians of the 50 replications in Table 2. The implications of these results are similar to those above. However, we see that the methods using the hyper-g prior also assign the highest probability to the true model if this is the null model. The difference to the methods using the hyper-g/n prior is that this highest probability is bounded away from unity, as we saw before.

Similarly to Ley and Steel (2012, section 7) we can also look at the posterior expectation of the model size, *i. e.* the posterior expected number of included covariates. Taking the median over the 50 replications, we get the results in Table 3. We observe that the parametric methods 3, 4 and 5 include too many covariates, especially in the null model case. Except for the large model with only $n = 50$, all other methods perform well.

In summary, the additive model selection procedures introduced in this paper are very competitive with the considered alternative methods concerning discovery of the

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	1	1	1	1	1	1	5	1	1
Hyper-g/n splines	1	1	1	1	1	1	7158	1	1
Hyper-g linear	1	1	1	618	62	6437	∞	2684	106270
Hyper-g/n linear	1	1	1	426	46	4564	∞	2356	76550
Bayesian FPs	1	1	1	6	1	4	∞	1	1
Spike-and-slab	1	1	1	3	1	1	∞	2	1
Knot selection	1	1	1	∞	1	1	∞	∞	1

Table 2 – *Median posterior rank of the true model, when the true model is defined by correct variable inclusion (for the MAP model the rank is 1). When the true model was not discovered at all by the method, the rank was recorded as ∞, leading to some infinite entries in this table.*

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	0.2	0.2	0.2	6.3	6.4	6.2	2.3	6.3	6.1
Hyper-g/n splines	0.2	0.1	0.0	6.4	6.4	6.1	48.0	6.3	6.1
Hyper-g linear	6.9	6.6	6.6	10.9	10.0	7.5	2.1	5.3	7.5
Hyper-g/n linear	1.9	1.0	0.2	8.2	8.3	7.2	48.0	5.1	7.2
Bayesian FPs	1.2	1.2	1.2	6.1	6.6	7.3	1.6	6.4	6.7
Spike-and-slab	0.1	0.1	0.0	5.7	6.6	6.2	98.1	6.5	6.1
Knot selection	0.1	0.1	0.0	4.4	6.8	6.1	3.1	6.1	6.1

Table 3 – *Median average posterior model size. The true numbers are 0, 6 and 6 for the null, small and large models, respectively.*

	small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	1 (0.67, 1)	1 (1, 1)	1 (1, 1)	0.17 (0, 1)	1 (1, 1)	1 (1, 1)
Hyper-g/n splines	1 (0.67, 1)	1 (1, 1)	1 (1, 1)	0.67 (0.33, 1)	1 (1, 1)	1 (1, 1)
Hyper-g linear	0.83 (0.41, 1)	0.83 (0.67, 1)	0.83 (0.83, 1)	0.17 (0, 0.33)	0.5 (0.33, 0.83)	0.83 (0.83, 0.83)
Hyper-g/n linear	0.67 (0.33, 1)	0.83 (0.5, 1)	0.83 (0.83, 0.83)	0.5 (0.07, 0.93)	0.5 (0.33, 0.83)	0.83 (0.83, 0.83)
Bayesian FPs	0.83 (0.33, 1)	1 (1, 1)	1 (1, 1)	0.17 (0, 0.77)	1 (0.83, 1)	1 (1, 1)
Spike-and-slab	0.75 (0.5, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	0.83 (0.67, 1)	1 (1, 1)
Knot selection	0.5 (0, 0.83)	1 (0.83, 1)	1 (1, 1)	0.17 (0, 0.6)	0.67 (0.41, 0.83)	1 (1, 1)

Table 4 *– Median sensitivity for variable inclusion based on 0.5 threshold on posterior inclusion probabilities, together with 5% and 95% quantiles.*

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (0.93, 1)	1 (0.93, 1)	1 (1, 1)	1 (0.99, 1)	1 (0.99, 1)	1 (1, 1)
Hyper-g/n splines	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (0.89, 1)	1 (0.93, 1)	1 (1, 1)	0.54 (0.51, 1)	1 (0.99, 1)	1 (1, 1)
Hyper-g linear	1 (0, 1)	1 (0.7, 1)	1 (0.72, 1)	0.82 (0, 1)	0.86 (0.23, 1)	0.93 (0.79, 0.93)	1 (0.99, 1)	1 (0.98, 1)	0.99 (0.98, 0.99)
Hyper-g/n linear	1 (0.64, 1)	1 (0.97, 1)	1 (1, 1)	0.93 (0.43, 1)	0.93 (0.68, 1)	0.93 (0.79, 0.93)	0.53 (0.51, 1)	1 (0.98, 1)	0.99 (0.98, 0.99)
Bayesian FPs	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (0.79, 1)	1 (0.93, 1)	0.93 (0.86, 1)	1 (0.99, 1)	1 (0.99, 1)	1 (0.98, 1)
Spike-and-slab	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (1, 1)	0 (0, 0)	1 (1, 1)	1 (1, 1)
Knot selection	1 (1, 1)	1 (1, 1)	1 (1, 1)	1 (0.86, 1)	1 (0.93, 1)	1 (1, 1)	1 (0.94, 1)	1 (0.98, 1)	1 (1, 1)

Table 5 *– Median specificity for variable inclusion based on 0.5 threshold on posterior inclusion probabilities, together with 5% and 95% quantiles.*

true set of influential covariates. In particular, they showed clear advantages in the case of small and moderate sample sizes. Using splines instead of only linear effects proved essential for the discovery of the masked effect of covariate x_4.

1.2.2 Variable inclusion

Each of the methods produces posterior inclusion probabilities for the considered covariates. Choosing the threshold 0.5 on these inclusion probabilities, we can compute the sensitivity (Table 4) and specificity for selecting the correct covariates (Table 5). The tables contain the median together with a 90% confidence interval obtained from the 50 replications. All methods except the linear only methods perform well, and the proposed hyper-g and hyper-g/n spline approaches are very competitive.

Moreover, we computed the area under the ROC curve (AUC) for variable inclusion

	small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	1.00	1.00	1.00	0.99	1.00	1.00
Hyper-g/n splines	1.00	1.00	1.00	0.69	1.00	1.00
Hyper-g linear	0.86	0.90	0.94	0.82	0.92	0.92
Hyper-g/n linear	0.86	0.90	0.94	0.67	0.92	0.92
Bayesian FPs	0.99	1.00	1.00	0.92	1.00	1.00
Spike-and-slab	1.00	1.00	1.00	0.43	1.00	1.00
Knot selection	0.88	1.00	1.00	0.81	0.95	1.00

Table 6 *– Median AUC for variable inclusion based on posterior inclusion probabilities.*

based on the posterior inclusion probabilities. The median AUC values over the 50 replications for the small and large model case are shown in Table 6. For the null model case the sensitivity and AUC could not be computed because no covariate had a true effect. Also in this table all methods, except the hyper-g and hyper-g/n linear approaches, perform well.

Of substantial interest is also if the methods can distinguish between the correct inclusion of covariates x_{16} and x_{17} and the wrong inclusion of the highly correlated covariates x_{18}, x_{19}, x_{20}. To this end we computed the difference of the corresponding average inclusion probabilities:

$$\frac{1}{2}(P_{16} + P_{17}) - \frac{1}{3}(P_{18} + P_{19} + P_{20})$$

where $P_j = \mathbb{P}\{m_j(x_j) \neq 0 \,|\, \boldsymbol{y}\}$, and averaged the difference over the 50 replications. The results are shown in Table 7. We see that the proposed spline methods work best, because already for $n = 50$ they can distinguish quite well.

Finally, in Table 8 we examine the difference of the inclusion probabilities for the surrogate x_8 and the truly influential covariate x_4. We see that with increasing sample size, the two linear methods choose x_8 instead of x_4. This result is analogous to the Pima Indian diabetes data example, where the strong nonlinear effect of x_7 was missed and instead x_1 got higher inclusion probabilities, when pure variable selection without co-

	small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	75	97	98	26	100	100
Hyper-g/n splines	79	97	98	20	100	100
Hyper-g linear	18	44	87	6	26	98
Hyper-g/n linear	22	48	90	17	26	98
Bayesian FPs	41	89	68	9	92	81
Spike-and-slab	30	88	97	1	60	97
Knot selection	9	78	99	4	13	99

Table 7 *– Average difference of inclusion probabilities (in percentage points) between the truly effective covariates x_{16} and x_{17} and the nuisance covariates x_{18}, x_{19}, x_{20}. (The optimal value is 100, the worst value is -100.)*

variate transformation was done. For the smallest sample size $n = 50$, the two proposed spline methods perform clearly better than the other approaches which are capable of fitting nonlinear effects. For the larger sample sizes $n = 100$ and $n = 1000$ this advantage is smaller.

Overall, the variable inclusion performance did not differ substantively with respect to sensitivity, specificity and AUC between the considered methods, with the exception of a slightly worse performance of the two linear methods. However, the hyper-g and hyper-g/n spline methods were clearly better in distinguishing truly effective covariates from highly correlated nuisance covariates. Moreover, for small sample sizes, they outperformed the other nonlinear methodologies concerning discovery of the masked x_4 effect. In this task the merely linear methods obviously failed.

1.2.3 Function estimation

For each covariate x_j, we can compute the mean squared error (MSE) of the model averaged function estimate $\hat{m}_j(x_j)$. (Note that we correctly account for the centring of the function estimates in this step.) Taking the mean over all covariates and the mean over all 50 replications, we obtain the average MSEs shown in Table 9. Considering the

	small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	89	96	99	52	99	100
Hyper-g/n splines	91	96	99	58	99	100
Hyper-g linear	-8	-20	-84	-2	-11	-95
Hyper-g/n linear	-8	-22	-87	4	-10	-95
Bayesian FPs	61	94	97	23	98	100
Spike-and-slab	79	95	96	6	97	90
Knot selection	52	91	100	27	89	100

Table 8 *– Average difference of inclusion probabilities (in percentage points) between the truly influential covariate x_4 and the surrogate x_8. (The optimal value is 100, the worst value is -100.)*

null model case, the spike-and-slab method produces the largest errors. In the small and large models, the Bayesian FPs produce the largest errors for $n = 50, 100$, while for $n = 1000$, the linear methods perform worst. Among the other methods, the knot selection approach is slightly inferior to the splines and spike-and-slab approaches.

We also investigated the coverage rates of pointwise 95% credible intervals for the functions. Taking the mean over all covariates and the mean over all 50 replications, we obtain the average coverage rates in Table 10. (Note that we compute means here because coverages are expected frequencies.) In the null model case, all approaches have very large coverage rates. In the small and large model cases, the two spline methods have slightly too large coverage compared to the nominal 95% level. The spike-and-slab method is even more conservative. The knot selection approach has good performance.

In summary, the proposed additive model selection procedures were very competitive concerning estimation of the partial linear predictor functions $m_j(x_j)$. While being slightly conservative in terms of coverage rates of credible intervals, they performed well or better than the best compared method in terms of MSEs. It is interesting that the hyper-g splines were slightly but consistently better than the hyper-g/n splines.

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	0.03	0.01	0.00	39.15	10.32	1.68	30.42	1.88	0.33
Hyper-g/n splines	0.05	0.01	0.00	47.82	18.33	3.20	784.44	2.78	0.61
Hyper-g linear	0.76	0.14	0.01	158.10	133.55	121.97	45.11	32.26	24.36
Hyper-g/n linear	0.22	0.02	0.00	189.57	169.00	120.96	378.07	36.23	26.09
Bayesian FPs	0.14	0.03	0.00	16837.92	3026.61	29.51	76.78	356.30	5.80
Spike-and-slab	1.90	1.82	0.57	80.94	14.00	2.09	45.45	8.71	0.81
Knot selection	0.03	0.00	0.00	180.03	35.29	2.07	47.23	29.33	0.78

Table 9 *– Average MSEs (in 10^{-4} units) of function estimates.*

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	100	100	100	96	97	95	97	99	99
Hyper-g/n splines	100	100	100	97	98	97	96	100	99
Hyper-g linear	100	100	100	87	86	78	95	97	96
Hyper-g/n linear	100	100	100	83	84	82	65	96	96
Bayesian FPs	100	100	100	91	93	88	96	99	98
Spike-and-slab	100	100	100	97	100	100	100	100	100
Knot selection	100	100	100	84	94	93	95	96	99

Table 10 *– Average coverage rates (percentages) of pointwise 95% credible intervals.*

	null			small			large		
	$n = 50$,	100,	1000	$n = 50$,	100,	1000	$n = 50$,	100,	1000
Hyper-g splines	3.5	3.3	19.3	9.1	6.8	14.4	48.2	9.4	22.7
Hyper-g/n splines	3.6	1.0	1.0	22.9	10.4	14.8	28.6	12.2	21.5
Hyper-g linear	0.7	0.8	2.1	0.6	0.5	0.4	3.7	3.7	3.5
Hyper-g/n linear	4.3	1.8	0.7	9.0	5.4	1.2	2.0	7.9	5.2
Bayesian FPs	9.2	9.7	27.6	18.3	20.2	27.4	20.9	37.9	69.1
Spike-and-slab	0.5	0.7	3.1	0.5	0.6	3.0	3.3	4.9	77.6
Knot selection	0.5	0.4	0.5	0.6	0.6	0.8	1.6	1.7	1.9

Table 11 *– Average computation times in minutes.*

1.2.4 Computational effort

Finally we report the average required time to complete the computations in Table 11. The timings were obtained on computing nodes of the supercomputer "Schroedinger" at the University of Zurich, having 2 quad-core processors (Intel, 2.8 GHz) and 24 GB RAM each. Note that 8 single-threaded simulations were run in parallel on each machine, so the reported computational effort per simulation is to be understood as non-parallelised and is expected on 2.8 GHz single-core CPUs.

Interestingly, the spline methods are overall slightly faster than the less flexible Bayesian FPs. The knot selection approach is the fastest method. Spike-and-slab is also fast, except for the large model with $n = 1000$, where it is much slower than, say, the spline approaches.

2 Additional example: Ozone data

We apply our additive modelling approach to the ozone data from Breiman and Friedman (1985) on the association between $p = 9$ meteorological covariates and the maximum one-hour average ozone concentration for $n = 330$ days in 1976 (see Table 12 for details). We use again cubic O'Sullivan splines (Wand and Ormerod, 2008). Here, we get basis matrices $\boldsymbol{Z}_j$ with $K = 6$ columns from 4 inner knots at the quintiles.

Variable	Description
y	Maximum 1-hour average ozone level [ppm]
x_1	Day of the year
x_2	500 millibar pressure height [m]
x_3	Wind speed [mph]
x_4	Relative humidity [%]
x_5	Temperature at Sandberg, CA [°F]
x_6	Inversion base height [feet]
x_7	Pressure gradient [mm Hg]
x_8	Visibility [miles]
x_9	Inversion base temperature [°F]

Table 12 – *Description of the variables in the ozone data set.*

Exhaustive evaluation of the posterior model probabilities $f(\boldsymbol{d} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{d}) f(\boldsymbol{d})$ of all $(K+1)^9 = 40\,353\,607$ models takes 531 minutes on a standard 2.8 GHz CPU. We applied the stochastic search algorithm with 10^6 iterations, which took 2 minutes and resulted in 125 619 models. 92.4% of the posterior model probability have been discovered and the 1858 top models were found by this algorithm.

In Table 13 the marginal posterior probabilities for linear and smooth inclusion of the nine covariates are shown. While x_1, x_5, x_7 and x_8 are clearly included as smooth terms, there is considerable uncertainty for the other covariates whether to be included linearly or smoothly. Only the overall inclusion probability for x_6 is below 50%.

The MAP model includes smooth terms for x_1, x_5, x_7, x_8 and x_9. The covariates x_2 and x_4 are included linearly while x_3 and x_6 are not included. Figure 2 shows the estimated covariate effects, which were obtained from 10 000 posterior samples. Note that for linear functions m_j, the pointwise credible intervals coincide with the simultaneous credible intervals (Besag, Green, Higdon, and Mengersen, 1995, p. 30). This is because all straight lines samples intersect in one point, which is $(\bar{x}_j, 0)$ due to the centring of the covariates.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
not included ($d_j = 0$)	0.00	0.02	0.33	0.06	0.00	0.59	0.00	0.00	0.28
linear ($d_j = 1$)	0.00	0.46	0.32	0.35	0.00	0.14	0.00	0.05	0.22
smooth ($d_j > 1$)	1.00	0.52	0.35	0.59	0.99	0.28	1.00	0.95	0.50

Table 13 *– Marginal posterior inclusion probabilities in the ozone data set.*

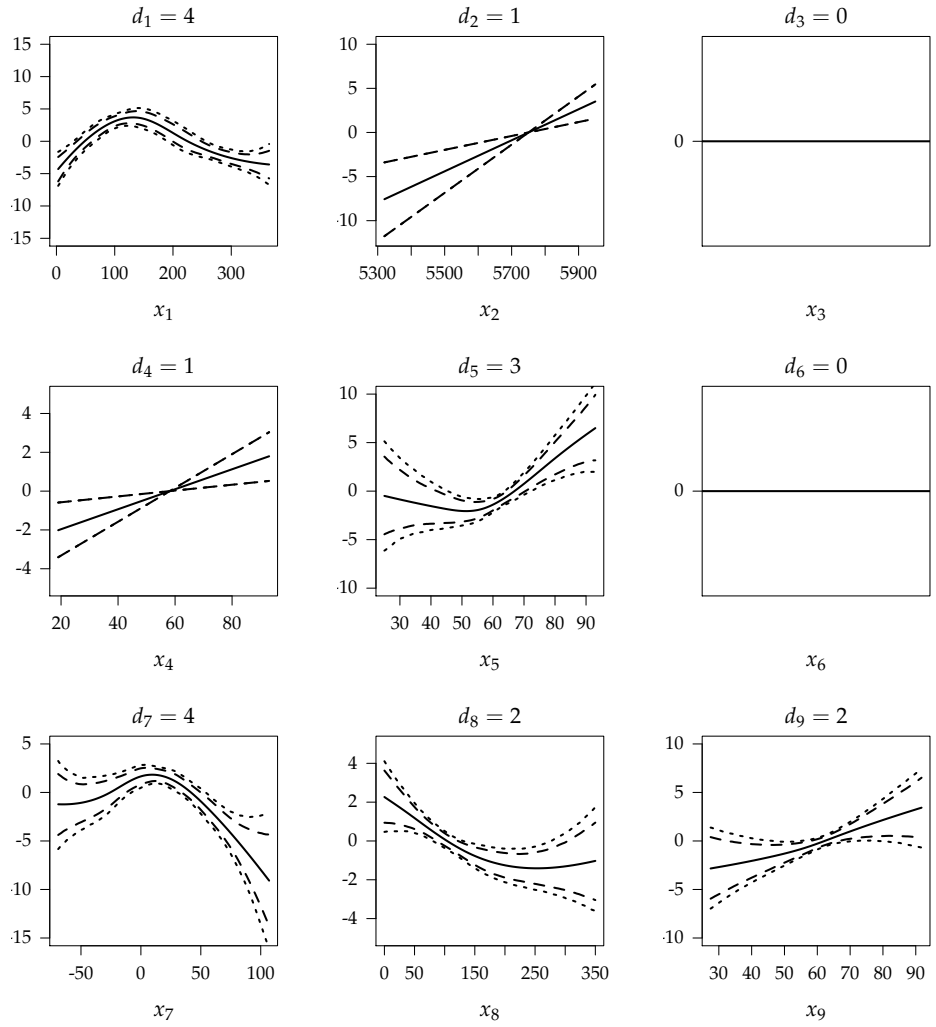


Figure 2 *– Estimated covariate effects in the MAP model for the ozone data, based on 10 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.*

In comparison to the MAP model in Sabanés Bové and Held (2011, section 4) based on Bayesian fractional polynomials, similar functional forms are estimated for the effects of x_1, x_4 and x_5, while differences are visible for x_7 and x_8. Note that x_6 is included in the MAP model in Sabanés Bové and Held (2011). See also Casella and Moreno (2006) for an objective Bayesian variable selection analysis (without the possibility of smooth effects) of this data set.

Given the list of all possible models $\boldsymbol{d} \in \mathcal{D}^p$, or a subset found by the stochastic search procedure, one may consider postprocessing the results.

The best meta-model for the ozone data features all covariates except x_6 and has posterior probability 0.261. The corresponding estimates of the covariate effects are shown in Figure 3. The best meta-model happens to be identical with the median probability meta-model, *cp*. Tables 13.

Second, in order to allow for continuous degrees of freedom, one can optimise the marginal likelihood of the MAP model with respect to the degrees of freedom of the covariates included. The MAP configuration for the ozone data is $(4, 1, 0, 1, 3, 0, 4, 2, 2)$ and the resulting optimised configuration is $(4.35, 1, 0, 1.08, 3.44, 0, 3.63, 2.3, 1)$, which increases the log marginal likelihood from -1413.86 to -1412.91. Although d_9 decreases from 2 to 1 (rounded down to 2 decimals), the function estimates are very similar to those from the MAP model in Figure 2.

3 Implementation details

Section 3.1 gives details on the efficient computation of the marginal likelihood for normal additive models. Section 3.2 gives details on the parameter sampling in a given normal additive model. Finally, Section 3.3 derives the proposal probabilities for the stochastic search procedure.

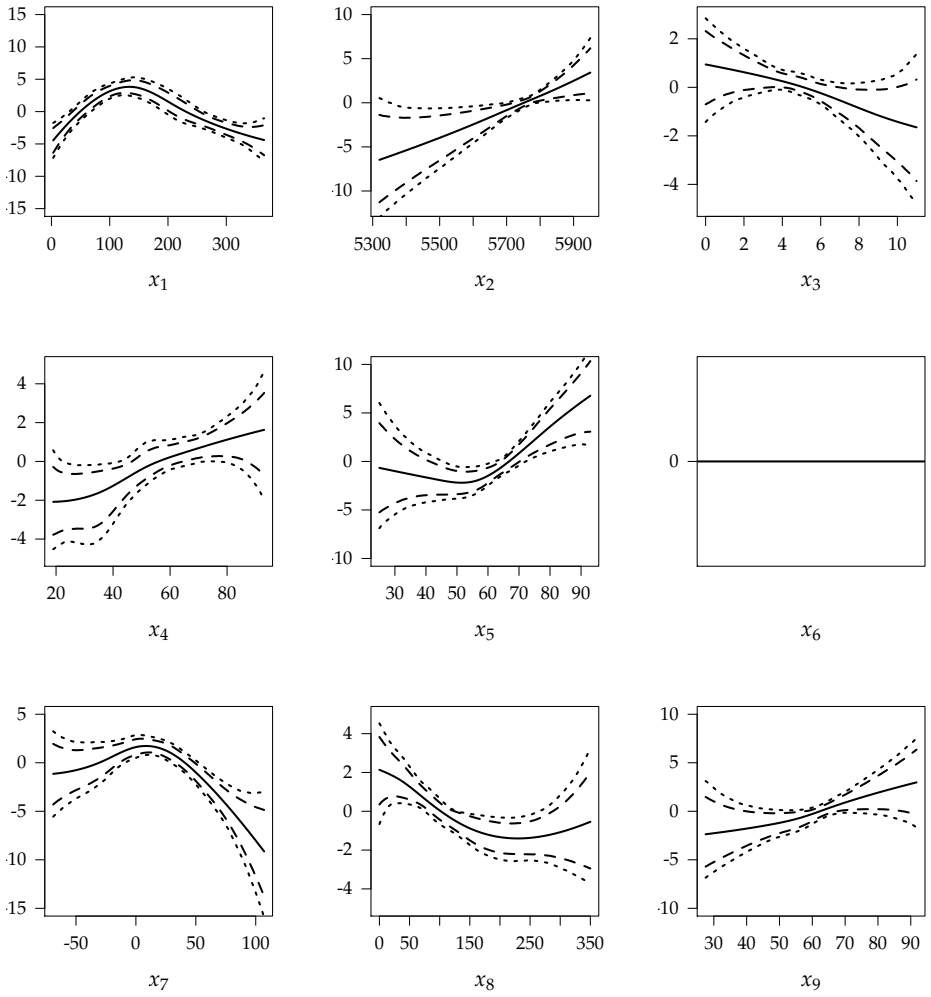


Figure 3 *– Estimated covariate effects in the best meta-model (and median probability meta-model) for the ozone data, based on 20 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals are shown.*

3.1 Marginal likelihood computation

For the coefficient of determination $\tilde{R}^2_d = SSM_d/SST_d$ required in the closed forms of the marginal likelihood for normal models (Appendix A), we need to compute the sum of squares in total (SST_d) and the sum of squares explained by the model (SSM_d). For SST_d, we have

$$\begin{aligned} SST_d &= (\boldsymbol{y} - \mathbf{1}_n\bar{y})^T \boldsymbol{V}_d^{-1}(\boldsymbol{y} - \mathbf{1}_n\bar{y}) \\ &= \|\boldsymbol{y} - \mathbf{1}_n\bar{y}\|^2 - \|\boldsymbol{W}_d^T(\boldsymbol{y} - \mathbf{1}_n\bar{y})\|^2. \end{aligned}$$

Note that the first term in the marginal likelihood under a hyper-g prior can be written as $\left\|\boldsymbol{V}_d^{-T/2}(\boldsymbol{y} - \mathbf{1}_n\bar{y})\right\|^{-(n-1)} = SST_d^{-(n-1)/2}$. For SSM_d, note that the fit of the general linear model is $\hat{\boldsymbol{y}}_d = \mathbf{1}_n\bar{y} + \boldsymbol{X}_d\hat{\boldsymbol{\beta}}_d$, where

$$\hat{\boldsymbol{\beta}}_d = (\boldsymbol{X}_d^T\boldsymbol{V}_d^{-1}\boldsymbol{X}_d)^{-1}\boldsymbol{X}_d^T\boldsymbol{V}_d^{-1}\boldsymbol{y}$$

is the weighted least squares estimate of $\boldsymbol{\beta}_d$. Therefore

$$\begin{aligned} SSM_d &= (\hat{\boldsymbol{y}}_d - \mathbf{1}_n\bar{y})^T \boldsymbol{V}_d^{-1}(\hat{\boldsymbol{y}}_d - \mathbf{1}_n\bar{y}) \\ &= \hat{\boldsymbol{\beta}}_d^T\boldsymbol{X}_d^T\boldsymbol{V}_d^{-1}\boldsymbol{X}_d\hat{\boldsymbol{\beta}}_d \end{aligned}$$

can be computed by Cholesky factorising $\boldsymbol{X}_d^T\boldsymbol{V}_d^{-1}\boldsymbol{X}_d = \boldsymbol{C}_d^T\boldsymbol{C}_d$, solving the triangular system $\boldsymbol{C}_d^T\boldsymbol{v}_d = \boldsymbol{X}_d^T\boldsymbol{V}_d^{-1}\boldsymbol{y}$ and setting $SSM_d = \|\boldsymbol{v}_d\|^2$.

For the computations above, we need the inverse of the covariance matrix $\boldsymbol{V}_d \in \mathbb{R}^{n\times n}$. While usually a Cholesky factorisation would be done, here it is advisable to avoid it because it has complexity $\mathcal{O}(n^3)$ and is therefore computationally expensive. Therefore, we instead work with the formula

$$\boldsymbol{V}_d^{-1} = \boldsymbol{I}_n - \boldsymbol{Z}_d\boldsymbol{M}_d^{-1}\boldsymbol{Z}_d^T$$

for the precision matrix, where $\boldsymbol{M}_d = \boldsymbol{Z}_d^T\boldsymbol{Z}_d + \boldsymbol{D}_d^{-1}$. The latter matrix has dimension JK, which is usually smaller than n, provided the spline basis dimension K is small. Thus, the Cholesky factorisation $\boldsymbol{M}_d = \boldsymbol{M}_d^{T/2}\boldsymbol{M}_d^{1/2}$ is relatively fast, and we compute $\boldsymbol{W}_d = \boldsymbol{Z}_d\boldsymbol{M}_d^{-1/2}$ such that $\boldsymbol{V}_d^{-1} = \boldsymbol{I}_n - \boldsymbol{W}_d\boldsymbol{W}_d^T$.

Finally, to compute the determinant term in the marginal likelihood which stems from the change of variables, we can again avoid factorising $\boldsymbol{V}_{\boldsymbol{d}}$, because we have

$$\left|\boldsymbol{V}_{\boldsymbol{d}}^{1/2}\right|^{-1} = \left|\boldsymbol{V}_{\boldsymbol{d}}^{-1}\right|^{1/2} = \left|\boldsymbol{I}_n - \boldsymbol{W}_{\boldsymbol{d}}\boldsymbol{W}_{\boldsymbol{d}}^T\right|^{1/2} = \left|\boldsymbol{I}_{JK} - \boldsymbol{W}_{\boldsymbol{d}}^T\boldsymbol{W}_{\boldsymbol{d}}\right|^{1/2},$$

see Harville (1997, p. 416) for the last equality. So again only a matrix of dimension JK, namely $\boldsymbol{I}_{JK} - \boldsymbol{W}_{\boldsymbol{d}}^T\boldsymbol{W}_{\boldsymbol{d}}$, needs to be factorised. Here, a LU factorisation can be used.

3.2 Parameter sampling

This section shall describe in detail the parameter sampling summarised in Section 2.1 of the paper.

Posterior inference in a given model $\boldsymbol{d}$ is based on Monte Carlo estimation of the parameters in the conditional model. We therefore use the factorisation

$$f(\beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \boldsymbol{u}_{\boldsymbol{d}}, \sigma^2, g \mid \boldsymbol{y}) = f(\boldsymbol{u}_{\boldsymbol{d}} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}, \sigma^2, \boldsymbol{y}) f(\beta_0, \boldsymbol{\beta}_{\boldsymbol{d}} \mid \sigma^2, g, \boldsymbol{y}) f(\sigma^2 \mid \boldsymbol{y}) f(g \mid \boldsymbol{y}). \quad (1)$$

Sampling of g, σ^2 and subsequently $\beta_0, \boldsymbol{\beta}_{\boldsymbol{d}}$ is done as follows.

Based on the decorrelated model, we first sample g. If the hyper-g prior is used, we perform inverse sampling: Let $\tilde{R}^2_{\boldsymbol{d}} = SSM_{\boldsymbol{d}}/SST_{\boldsymbol{d}}$ be again the coefficient of determination, and let $a_g = (n-p-3)/2$ and $b_g = (p+2)/2$ be the parameters of a beta distribution with cumulative distribution function $B_{\boldsymbol{d}}$, set

$$h = \frac{1 - \tilde{R}^2_{\boldsymbol{d}}}{B_{\boldsymbol{d}}^{-1}\{u + (1-u)B_{\boldsymbol{d}}(1 - \tilde{R}^2_{\boldsymbol{d}})\}}$$

where $u \sim \mathrm{U}(0,1)$, then $g = (1-h)/\tilde{R}^2_{\boldsymbol{d}}$ is a sample from $f(g \mid \boldsymbol{y})$. If the hyper-g/n prior is used, a numerical approximation of the quantile function can be used. First, the unnormalised log posterior density function $\log\{\tilde{f}(z, \boldsymbol{y} \mid \boldsymbol{d})\}$ of $z = \log(g/n)$ is maximised at $\hat{z}$. Then the `R`-package `Runuran` (see Leydold and Hörmann, 2009) can be used (function `pinv.new`) to generate a random sampler for the corresponding distribution. Finally the sampled z is transformed to $g = \exp(z)n/\{1 + \exp(z)n\}$.

Second, σ^2 is sampled from an inverse-gamma distribution, with parameters $a_{\sigma^2} = (n-1)/2$ and $b_{\sigma^2} = (1 - t\tilde{R}^2_{\boldsymbol{d}})SST_{\boldsymbol{d}}/2$, where $t = g/(1+g)$ is the shrinkage factor.

Third, the intercept β_0 is sampled from a univariate normal distribution with mean $\bar{y}$ and variance σ^2/n. Then the covariate effects vector $\boldsymbol{\beta_d}$ is sampled from a multivariate normal distribution with mean $t\hat{\boldsymbol{\beta}}_{\boldsymbol{d}}$ and covariance matrix $t\sigma^2(\tilde{\boldsymbol{X}}_{\boldsymbol{d}}^T\tilde{\boldsymbol{X}}_{\boldsymbol{d}})^{-1}$.

Finally, the spline coefficient vector $\boldsymbol{u_d}$ is sampled from

$$\begin{aligned} f(\boldsymbol{u_d} \mid \beta_0, \boldsymbol{\beta_d}, \sigma^2, \boldsymbol{y}) &\propto f(\boldsymbol{u_d} \mid \sigma^2) f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta_d}, \boldsymbol{u_d}, \sigma^2) \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2}\left[\boldsymbol{u_d}^T \boldsymbol{D_d}^{-1} \boldsymbol{u_d} + \|\boldsymbol{y} - \mathbf{1}_n\beta_0 - \boldsymbol{X_d}\boldsymbol{\beta_d} - \boldsymbol{Z_d}\boldsymbol{u_d}\|^2 \right] \right\} \\ &\propto \mathrm{N}_{JK}\left(\boldsymbol{u_d} \mid \boldsymbol{\Sigma_d}\boldsymbol{Z_d}^T(\boldsymbol{y} - \boldsymbol{X_d}\boldsymbol{\beta_d}),\, \sigma^2\boldsymbol{\Sigma_d} \right), \end{aligned} \tag{2}$$

where $\boldsymbol{\Sigma_d} = (\boldsymbol{Z_d}^T\boldsymbol{Z_d} + \boldsymbol{D_d}^{-1})^{-1}$ and β_0 disappears because $\boldsymbol{Z_d}^T\mathbf{1}_n = \mathbf{0}_{JK}$.

Given posterior samples for the linear coefficient β_j and the spline coefficient vector $\boldsymbol{u_j}$ for covariate j ($d_j > 1$), we would like to transform these into samples for the function $m_j(x_j)$, along a grid vector $\tilde{\boldsymbol{x}}_j^*$ of n^* points (on the same scale as the original $\tilde{\boldsymbol{x}}_j$ used for the model fitting). This is in principle straightforward, but one has to carefully apply analogous transformations as in (4) and (5) to $\tilde{\boldsymbol{x}}_j^*$ and the corresponding spline basis matrix $\tilde{\boldsymbol{Z}}_j^*$:

$$\boldsymbol{x}_j^* = \tilde{\boldsymbol{x}}_j^* - \mathbf{1}_{n^*}\frac{\mathbf{1}_n^T\tilde{\boldsymbol{x}}_j}{\mathbf{1}_n^T\mathbf{1}_n}, \tag{3}$$

$$\boldsymbol{Z}_j^* = \tilde{\boldsymbol{Z}}_j^* - \mathbf{1}_{n^*}\frac{\mathbf{1}_n^T\tilde{\boldsymbol{Z}}_j}{\mathbf{1}_n^T\mathbf{1}_n} - \boldsymbol{x}_j^*\frac{\boldsymbol{x}_j^T\tilde{\boldsymbol{Z}}_j}{\boldsymbol{x}_j^T\boldsymbol{x}_j}. \tag{4}$$

Afterwards, for each coefficient sample one can compute the corresponding vector of function values $m_j(\tilde{\boldsymbol{x}}_j^*) = \boldsymbol{x}_j^*\beta_j + \boldsymbol{Z}_j^*\boldsymbol{u}_j$. Similarly, prediction samples for the corresponding response vector $\boldsymbol{y}^*$ can be extracted from the sampling output.

3.3 Proposal probabilities

First note that the two proposal types 'Move' and 'Swap' do not overlap, because a 'Move' always changes exactly one d_j, while a 'Swap' either changes none or two d_j's. Denote with p_m the probability to choose a 'Move'.

Suppose a 'Move' was proposed for covariate $j \in \{0, 1, \ldots, p\}$. We then have

$$q(\boldsymbol{d}' \mid \boldsymbol{d}) = p_m \cdot \frac{1}{p} \cdot \begin{cases} 1, & d_j \in \{0, K\}, \\ \frac{1}{2}, & \text{else} \end{cases}$$

and analogously

$$q(\boldsymbol{d} \mid \boldsymbol{d}') = p_m \cdot \frac{1}{p} \cdot \begin{cases} 1, & d'_j \in \{0, K\}, \\ \frac{1}{2}, & \text{else} \end{cases}$$

with proposal ratio

$$\frac{q(\boldsymbol{d}' \mid \boldsymbol{d})}{q(\boldsymbol{d} \mid \boldsymbol{d}')} = \begin{cases} \frac{1}{2}, & d'_j \in \{0, K\}, \\ 2, & d_j \in \{0, K\}, \\ 1, & \text{else.} \end{cases}$$

For the 'Swap' proposal, suppose covariates i and j are proposed to interchange their model parameters d_i and d_j. Of course, if $d_i = d_j$, then the proposal ratio equals unity because $\boldsymbol{d}' = \boldsymbol{d}$. In the other case, both model parameters are changed, and

$$q(\boldsymbol{d}' \mid \boldsymbol{d}) = q(\boldsymbol{d} \mid \boldsymbol{d}') = (1 - p_m) \cdot \binom{p}{2}^{-1},$$

so that for a 'Swap' we always have $q(\boldsymbol{d}' \mid \boldsymbol{d}) / q(\boldsymbol{d} \mid \boldsymbol{d}') = 1$.

References

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian computation and stochastic systems (with discussion)," *Statistical Science*, 10, 3–66.

Breiman, L. and Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, 80, 580–598.

Casella, G. and Moreno, E. (2006), "Objective Bayesian variable selection," *Journal of the American Statistical Association*, 101, 157–167.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Wiley Series in Probability and Statistics, Chichester: Wiley.

Harville, D. A. (1997), *Matrix Algebra From a Statistician's Perspective*, New York: Springer.

Ley, E. and Steel, M. F. (2012), "Mixtures of g-priors for Bayesian model averaging with economic applications," *Journal of Econometrics*, 171, 251–266.

Leydold, J. and Hörmann, W. (2009), *UNU.RAN: A Library for Non-Uniform Universal Random Variate Generation, Version 1.4.*, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, available at http://statmath.wu.ac.at/software/unuran/index.html.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of *g* priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

Sabanés Bové, D. and Held, L. (2011), "Bayesian fractional polynomials," *Statistics and Computing*, 21, 309–324.

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012), "Spike-and-slab priors for function selection in structured additive regression models," *Journal of the American Statistical Association*, 107, 1518–1532.

Wand, M. P. and Ormerod, J. T. (2008), "On semiparametric regression with O'Sullivan penalized splines," *Australian & New Zealand Journal of Statistics*, 50, 179–198.

PAPER III

Comment on Cai and Betensky (2003), On the Poisson approximation for hazard regression

Daniel Sabanés Bové & Leonhard Held

Letter to the Editor published in *Biometrics*, 2013, **69**, 795.

Biometrics 69, 795–796
September 2013

DOI: 10.1111/biom.12088

CORRESPONDENCE

Comment on Cai and Betensky (2003), On the Poisson Approximation for Hazard Regression

from: *Daniel Sabanés Bové* and Leonhard Held*
Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, 8001 Zurich, Switzerland
**email: daniel.sabanesbove@ifspm.uzh.ch*

Cai and Betensky (2003, Section 5.1) consider hazard regression for right-censored survival data. For any log-linear parametrization of the baseline hazard, they provide a Poisson approximation of the proportional hazards model. However, the derivation of their approximation is flawed, resulting in substantial bias that does not vanish with increasing sample size n. Indeed, the difference to the log-likelihood from the correct and consistent Poisson approximation given below is of order $\mathcal{O}(n)$. Hence the Cai and Betensky (2003) approximation produces completely different estimates than the exact likelihood.

To spot the error, consider ordered survival times t_i with censoring indicators δ_i, $i = 1, \ldots, n$. Under the proportional hazards assumption, the cumulative hazard $\Lambda_i(t)$ for the ith individual factors into the cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$ and the contribution $\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ of the covariates $\boldsymbol{x}_i$, $\Lambda_i(t) = \Lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. If we write $\lambda_0(t) = \exp\{\boldsymbol{z}(t)^T\boldsymbol{\gamma}\}$ in terms of a basis $\boldsymbol{z}(t)$, the log-likelihood is

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \delta_i \{\boldsymbol{z}(t_i)^T\boldsymbol{\gamma} + \boldsymbol{x}_i^T\boldsymbol{\beta}\} - \sum_{i=1}^{n} \Lambda_i(t_i). \qquad (1)$$

Using the trapezoidal cubature approximation $\Lambda_0(t_i) \approx \sum_{j=1}^{i} q_{ij}\lambda_0(t_j)$ from Cai, Hyndman, and Wand (2002, Section 4), where the q_{ij}'s are the entries of a lower-triangular matrix $\boldsymbol{Q}$, we obtain

$$\sum_{i=1}^{n} \Lambda_i(t_i) \approx \sum_{i=1}^{n} \sum_{j=1}^{i} q_{ij}\lambda_0(t_j)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}). \qquad (2)$$

In contrast, Cai and Betensky (2003, Section 5.1, third equation) give an approximation which translates to replacing $\boldsymbol{x}_i$ with $\boldsymbol{x}_j$ in (2), leading to their Poisson approximation

$$\delta_i \sim \text{Poisson}\left[\exp\left\{\log\left(\textstyle\sum_k q_{ki}\right) + \boldsymbol{z}(t_i)^T\boldsymbol{\gamma} + \boldsymbol{x}_i^T\boldsymbol{\beta}\right\}\right],$$

$i = 1, \ldots, n$. To obtain the correct Poisson approximation, plug (2) into (1), which yields the log-likelihood of

$$y_{ij} \sim \text{Poisson}\left[\exp\left\{\log(q_{ij}) + \boldsymbol{z}(t_j)^T\boldsymbol{\gamma} + \boldsymbol{x}_i^T\boldsymbol{\beta}\right\}\right],$$

$i = 1, \ldots, n$ and $j = 1, \ldots, i$, where the response is $y_{ij} = \delta_i$ for $i = j$ and $y_{ij} = 0$ otherwise and $\log(q_{ij})$ is a fixed offset.

`R` example code is available from the authors.

References

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**, 570–579.

Cai, T., Hyndman, R. J., and Wand, M. P. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics* 11, 784–798.

The authors replied as follows:

The main focus of Cai and Betensky (2003) was for the analysis of interval censored data under the Cox model. They proposed to weakly parameterize the baseline hazard using a piecewise linear spline and maximize the likelihood function with a PQL approximation. Theoretical and numerical results demonstrated the validity of the proposed point and interval estimators for the general case. For the special case with right censored data, Cai and Betensky (2003) suggested the use of Poisson mixed model to approximate the likelihood for computational ease. However, as pointed out by Bové and Held (2013), the Poisson approximation given in Cai and Betensky (2003) was incorrect. Bové and Held (2013) suggested to create pseudo observations $y_{ij} = I(i = j)\delta_i$ and fit a Poisson mixed model to $\{y_{ij} : i = 1, \ldots, n, j = 1, \ldots, i\}$. This approach will indeed lead to a valid approximation for the point estimator. On the other hand, since their new proposal involves a pseudo data with $n(n+1)/2$ observations, there might be some computational burden in fitting such a Poisson mixed model even with moderate sample sizes such as $n = 500$. Furthermore, one may not directly obtain correct standard error estimates from the fitting since the y_{ij}'s are not independent observations.

An alternative strategy to obtain the maximizer of

$$\ell(\boldsymbol{\theta}; \sigma_b) = \boldsymbol{\delta}^{\mathrm{T}}(\mathbf{X}_-^{\mathrm{T}}\boldsymbol{\theta}_- + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}) - \sum_{i=1}^{n} \Lambda_i(t_i) - \frac{1}{2\sigma_b^2}\mathbf{b}^{\mathrm{T}}\mathbf{b} - \frac{K}{2}\sigma_b^2$$

is to consider an iterative procedure based on

$$\begin{aligned}\sum_{i=1}^{n} \Lambda_i(t_i) &\approx \sum_{i=1}^{n} \sum_{j=1}^{i} q_{ij}\lambda_0(T_j)\exp(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{i} \exp\{\eta_0(T_j)\} q_{ij}\exp(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta})\end{aligned}$$

using the same notation as Bové and Held, where $\boldsymbol{\theta}_- = (\alpha_0, \alpha_1, \mathbf{b}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{X}_- = [1, T_i, (T_i - \boldsymbol{\kappa})_+]_{i=1,\ldots,n}$. Specifically, obtain an initial estimate of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}^{(0)}$, say as the standard max-

imum partial likelihood estimator. Then one may iterate via the following steps starting from $m = 1$:

(1) Let $q_{ij}^{(m\star)} = q_{ij} \exp(\mathbf{Z}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^{(m-1)})$ and $\bar{\mathbf{q}}^{(m\star)} = [\sum_{i=j}^{n} q_{ij}^{(m\star)}]_{j=1,\ldots,n}$. Maximize

$$\ell_p^{(m\star)}(\boldsymbol{\theta}_-; \sigma_b) \approx \boldsymbol{\delta}^{\mathrm{T}} \mathbf{X}_-^{\mathrm{T}} \boldsymbol{\theta}_- - \mathbf{1}^{\mathrm{T}} \exp\left\{\mathbf{X}_-^{\mathrm{T}} \boldsymbol{\theta}_- + \log(\bar{\mathbf{q}}^{(m\star)})\right\} - \frac{1}{2\sigma_b^2} \mathbf{b}^{\mathrm{T}} \mathbf{b} - \frac{K}{2} \sigma_b^2$$

with respect to $\{\boldsymbol{\theta}_-, \sigma_b\}$ to obtain $\{\widehat{\boldsymbol{\theta}}_-^{(m)}, \widehat{\sigma}_b^{(m)}\}$ and the corresponding $\widehat{\eta}_0^{(m)}(T_j)$.

(2) Let $q_{ij}^{(m\dagger)} = q_{ij} \exp\{\widehat{\eta}_0^{(m)}(T_j)\}$ and $\bar{\mathbf{q}}^{(m\dagger)} = [\sum_{j=1}^{i} q_{ij}^{(m\dagger)}]_{i=1,\ldots,n}$. Maximize

$$\ell_p^{(m\dagger)}(\boldsymbol{\beta}) \approx \boldsymbol{\delta}^{\mathrm{T}} \mathbf{Z}^{\mathrm{T}} \boldsymbol{\beta} - \mathbf{1}^{\mathrm{T}} \exp\left\{\mathbf{Z}^{\mathrm{T}} \boldsymbol{\beta} + \log(\bar{\mathbf{q}}^{(m\dagger)})\right\}$$

with respect to $\boldsymbol{\beta}$ to obtain $\widehat{\boldsymbol{\beta}}^{(m)}$.

(3) Let $m = m + 1$ and go back to Step (1) until convergence.

The maximizations can be achieved by fitting a Poisson mixed model with offset in Step (1) and a standard Poisson model with offset in Step (2).

Tianxi Cai[1,*] *and Rebecca Betensky*[2]
[1] *Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.*
[2] *Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.*
[*] *email: tcai@hsph.harvard.edu*

Biometrics 000, 000–000
000 0000

DOI: 000

Extended version of "Comment on Cai and Betensky (2003), On the Poisson approximation for hazard regression" by Daniel Sabanés Bové and Leonhard Held

Daniel Sabanés Bové* and Leonhard Held**

Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich,
8001 Zurich, Switzerland
**email:* daniel.sabanesbove@ifspm.uzh.ch
***email:* leonhard.held@ifspm.uzh.ch

1. Trapezoidal cubature approximation

Consider right-censored survival data with ordered survival times t_i and censoring indicators δ_i, $i = 1, \ldots, n$. If $\delta_i = 1$ the death time has been observed while for $\delta_i = 0$ the observation is censored, i.e. it is only known that death happened at a time larger than t_i. Cai et al. (2002) consider hazard estimation without additional covariates, and parametrize the baseline hazard as $\lambda_0(t) = \exp\{\boldsymbol{z}(t)^T\boldsymbol{\gamma}\}$ using a basis $\boldsymbol{z}(t)$. Then the log-likelihood for the coefficient vector $\boldsymbol{\gamma}$ can be written as

$$l(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \delta_i \boldsymbol{z}(t_i)^T\boldsymbol{\gamma} - \sum_{i=1}^{n} \Lambda_0(t_i), \tag{1}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$ is the cumulative baseline hazard. Cai et al. (2002, section 4) propose to use trapezoidal cubature of $\lambda_0(t)$ with knots at the survival times t_i to approximate $\Lambda_0(t)$. Since the cubature formula is linear in the integrand function values $\lambda_0(t_i)$ and the survival times are ordered increasingly, there is a lower-triangular matrix $\boldsymbol{Q} = (q_{ij})$ such that $\Lambda_0(t_i) \approx \sum_{j=1}^{n} q_{ij}\lambda_0(t_j)$, which leads to the following approximation of the second sum in (1):

$$\begin{aligned}
\sum_{i=1}^{n} \Lambda_0(t_i) &\approx \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}\lambda_0(t_j) \\
&= \sum_{j=1}^{n} \lambda_0(t_j) \sum_{i=1}^{n} q_{ij} \\
&= \sum_{j=1}^{n} \exp\{\boldsymbol{z}(t_j)^T\boldsymbol{\gamma} + \log(\textstyle\sum_k q_{kj})\}.
\end{aligned}$$

Hence the log-likelihood (1) is approximated by

$$l(\boldsymbol{\gamma}) \approx \sum_{i=1}^{n} \delta_i \boldsymbol{z}(t_i)^T\boldsymbol{\gamma} - \exp\{\boldsymbol{z}(t_i)^T\boldsymbol{\gamma} + \log(\textstyle\sum_k q_{ki})\},$$

which corresponds to the log-likelihood of a log-linear Poisson model with response δ_i, covariates $\boldsymbol{z}(t_i)$ and offset $\log(\sum_k q_{ki})$, $i = 1, \ldots, n$.

In their trapezoidal cubature approximation, Cai et al. (2002) implicitly assume a constant baseline hazard function on $[0, t_1]$, which leads to $\Lambda_0(t_1) \approx \lambda_0(t_1)t_1$. We can slightly improve their approximation by also applying the trapezoidal rule to this first interval:

$$\Lambda_0(t_1) \approx \frac{1}{2}\{\lambda_0(0) + \lambda_0(t_1)\}t_1.$$

We achieve this by including an additional pseudo-observation $t_0 = 0, \delta_0 = 0$ in the data set. Clearly $\Lambda_0(0) = 0$, so the first row of the enlarged $(n+1) \times (n+1)$ matrix $\boldsymbol{Q} = (q_{ij})$ will contain only zeroes. The whole matrix has the following form:

$$\boldsymbol{Q} = \frac{1}{2}\begin{pmatrix}
0 & 0 & \cdots & & & & 0 \\
t_1 & t_1 - t_0 & 0 & \cdots & & & 0 \\
t_1 & t_2 - t_0 & t_2 - t_1 & 0 & \cdots & & 0 \\
t_1 & t_2 - t_0 & t_3 - t_1 & t_3 - t_2 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \ddots & \ddots & \vdots \\
 & & & & & & 0 \\
t_1 & t_2 - t_0 & t_3 - t_1 & t_4 - t_2 & \cdots & & t_n - t_{n-1}
\end{pmatrix}$$

Note that below the main diagonal of $\boldsymbol{Q}$ the entries in the columns are constant. Therefore it is sufficient to define the main diagonal entries and the first off-diagonal entries. The pseudo-code in Algorithm 1 does this and also accommodates the case of ties between the survival times t_i. Cai et al. (2002) give another modification of $\boldsymbol{Q}$ for tied survival times, but our definition has the advantage that it can also be applied in the proportional hazard regression case in Section 2.

2. Hazard regression

Cai and Betensky (2003, section 5.1) include additional covariate values $\boldsymbol{x}_i$ in the regression model, leading to the hazard $\lambda_i(t) = \lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ and cumulative hazard $\Lambda_i(t) = \Lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ for the ith individual. The log-likelihood of all regression coefficients is hence

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=0}^{n} \delta_i\{\boldsymbol{z}(t_i)^T\boldsymbol{\gamma} + \boldsymbol{x}_i^T\boldsymbol{\beta}\} - \sum_{i=0}^{n} \Lambda_i(t_i). \tag{2}$$

Cai and Betensky (2003, section 5.1, third equation) give the formula "$\mathbf{1}^T \exp(\boldsymbol{X}^r\boldsymbol{\theta} + \boldsymbol{o})$" as an approximation to the

This document has been typeset in *Biometrics* style

Algorithm 1: Computation of $\boldsymbol{Q}$ entries

Input: Survival times $0 = t_0 < t_1 \leq t_2 \leq \cdots \leq t_n$.
Output: Main diagonal entries q_{ii}, $i = 0, \ldots, n$, and adjacent lower diagonal entries $q_{i-1,i}$, $i = 1, \ldots, n$, of the matrix $\boldsymbol{Q} = (q_{ij})_{0 \leq i,j \leq n}$.

```
Set q_{0,0} ← 0;
for i ← 1 to n do
    Δt ← t_i − t_{i−1};
    if Δt > 0 then
        q_{i,i} ← ½Δt;
        q_{i,i−1} ← q_{i−1,i−1} + q_{i,i};
        k ← i − 1;
    else
        q_{i,i} ← ½(t_i − t_k);
        q_{i,i−1} ← 0;
```

second sum in (2). This translates to

$$\sum_{j=0}^{n} \Lambda_j(t_j) \approx \sum_{j=0}^{n} \exp\{\boldsymbol{z}(t_j)^T\boldsymbol{\gamma} + \boldsymbol{x}_j^T\boldsymbol{\beta} + \log(\textstyle\sum_k q_{kj})\} = \sum_{i=0}^{n}\sum_{j=0}^{i} q_{ij}\lambda_0(t_j)\exp(\boldsymbol{x}_j^T\boldsymbol{\beta}). \quad \text{(A1)}$$

Cai and Betensky (2003) give trapezoidal approximation of the cumulative baseline hazards as rationale. However, from this trapezoidal approximation $\Lambda_0(t_i) \approx \sum_{j=0}^{n} q_{ij}\lambda_0(t_j)$ we actually obtain

$$\sum_{i=0}^{n} \Lambda_i(t_i) = \sum_{i=0}^{n} \Lambda_0(t_i)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \approx \sum_{i=0}^{n}\sum_{j=0}^{i} q_{ij}\lambda_0(t_j)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}). \quad \text{(A2)}$$

The difference is that $\boldsymbol{x}_i$ in (A2) is replaced by $\boldsymbol{x}_j$ in (A1). The substantial bias of the approximation (A1) resulting from this error is illustrated in Section 3.

Fortunately, we can still rewrite the approximate log-likelihood as that arising from a log-linear Poisson model, following the tradition of Holford (1980) and Aitkin and Clayton (1980). However, the required pseudo data set is of size $\mathcal{O}(n^2)$, and not $\mathcal{O}(n)$ as promised by Cai and Betensky (2003). We can rewrite the first sum in the log-likelihood (2):

$$\sum_{i=0}^{n} \delta_i\{\boldsymbol{z}(t_i)^T\boldsymbol{\gamma} + \boldsymbol{x}_i^T\boldsymbol{\beta}\} = \sum_{i=0}^{n}\sum_{j=0}^{i} y_{ij}\{\boldsymbol{z}(t_j)^T\boldsymbol{\gamma} + \boldsymbol{x}_i^T\boldsymbol{\beta}\} \quad (3)$$

with $y_{ii} = \delta_i$ and $y_{ij} = 0$ for all $i \neq j$. Combining (3) with the approximation (A2), we see that we can approximate the log-likelihood (2) with that of a log-linear Poisson model with $(n+1)(n+2)/2$ pseudo-observations y_{ij} as defined above, linear predictors $\boldsymbol{z}(t_j)^T\boldsymbol{\gamma} + \boldsymbol{x}_i^T\boldsymbol{\beta}$ and offsets $\log(q_{ij})$, $j = 0, \ldots, i$ and $i = 0, \ldots, n$. If there are ties in the observed survival times, such that $t_i = t_{i+1}$, then the observations y_{ki} with $k > i$ are not included in the pseudo data set, because $q_{ki} = 0$ in that case (compare Algorithm 1).

3. Example

We illustrate the substantial bias of the approximation (A1) from Cai and Betensky (2003) with data from a clinical trial reported in Embury et al. (1977) on the efficacy of maintenance chemotherapy for acute myelogenous leukemia, see also Miller (1981). 11 patients received treatment ($x = 1$), while 12 patients did not ($x = 0$). We only include the treatment x as a covariate in the analysis, such that $\lambda_i(t) = \lambda_0(t)\exp(x_i\beta)$. We parametrize the baseline hazard as $\lambda_0(t) = \exp(\gamma_0 + \gamma_1 t)$. In this special case the cumulative baseline hazard is analytically available:

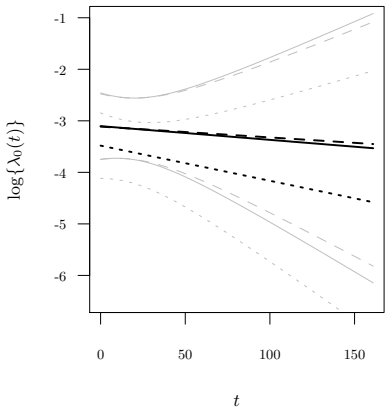
$$\Lambda_0(t) = \begin{cases} \exp(\gamma_0)t, & \text{if } \gamma_1 = 0, \\ \exp(\gamma_0)/\gamma_1\{\exp(\gamma_1 t) - 1\}, & \text{else,} \end{cases}$$

so the log-likelihood (2) can be computed analytically. In general this is not possible, and numerical integration methods would have to be used to obtain more accurate results than with the proposed Poisson approximation.

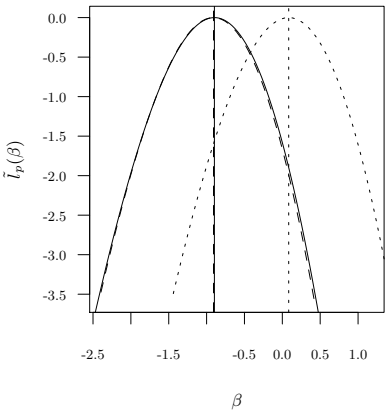
In Figure 1 we compare the results from maximum likelihood estimation under the exact (E) log-likelihood (2) and the Poisson model approximations using either the formula (A1) from Cai and Betensky (2003) or the corrected formula (A2). The log baseline hazard function estimates in Figure 1a are very similar for E and A2, while A1 is shifted. Also the relative profile log-likelihood functions for the covariate coefficient β in Figure 1b are very similar for E and A2, whereas A1 is shifted to the right. The MLEs (95% profile likelihood confidence intervals) for β are -0.897 (-1.985, 0.09) for E, -0.909 (-1.981, 0.067) for A2, and 0.084 (-1.013, 1.085) for A1. This illustrates that the resulting inference for the treatment effect differs greatly between E/A2 and A1: While the latter would conclude no noticeable treatment effect, the former suggest a protecting effect of the treatment. Standard Cox regression (Cox, 1972) gives the estimate -0.922 with 95% Wald confidence interval (-1.934, 0.09), which is close to E and A2.

We now illustrate that the bias of the approximation A1 does not vanish with increasing sample size. To this end, we simulate a much larger data set of size $n = 1000$, which resembles the original data from Embury et al. (1977), as follows: We draw 500 survival times in the treatment group uniformly in the range $[9, 161]$ of the survival times in the original treatment group. Similarly, 500 survival times have been generated in the control group, by sampling uniformly in the range $[5, 45]$ of the survival times in the original control group. The censoring indicators are set to 1 with probability 0.78, which is the fraction of observed survival times in the original data set.

In Figure 2 we compare again the results from the three different fitting methods. The log baseline hazard function estimates in Figure 2a of E and A2 overlap, while A1 is shifted downwards. Also the relative profile log-likelihood functions for the covariate coefficient β in Figure 2b overlap between E and A2, while A1 is substantially shifted to the right. The MLEs (95% profile likelihood confidence intervals) for β are -2.405 (-2.63, -2.191) for E, so only slightly different from -2.407 (-2.629, -2.19) obtained using A2. The difference to approximation A1 with MLE -0.727 (-0.955, -0.503) is substantial.

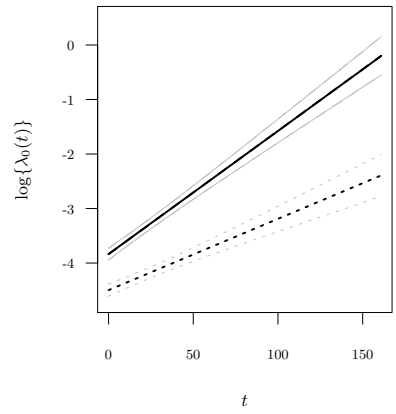


(a) Estimated log baseline hazard functions (black lines) with pointwise 95% confidence intervals (grey lines).

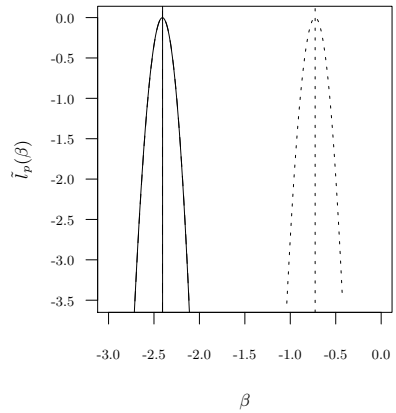


(b) Relative profile log-likelihood functions $\tilde{l}_p(\beta) = \max_{\boldsymbol{\gamma}} l(\beta, \boldsymbol{\gamma}) - l(\widehat{\beta}_{\mathrm{ML}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{ML}})$ and MLEs $\widehat{\beta}_{\mathrm{ML}}$.

Figure 1: Results from the exact computation (continuous lines) and the Poisson models with approximations (A2) (dashed lines) and (A1) (dotted lines), respectively.



(a) Estimated log baseline hazard functions (black lines) with pointwise 95% confidence intervals (grey lines).



(b) Relative profile log-likelihood functions $\tilde{l}_p(\beta) = \max_{\boldsymbol{\gamma}} l(\beta, \boldsymbol{\gamma}) - l(\widehat{\beta}_{\mathrm{ML}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{ML}})$ and MLEs $\widehat{\beta}_{\mathrm{ML}}$.

Figure 2: Simulated large data set: Results from the exact computation (continuous lines) and the Poisson models with approximations (A2) (dashed lines) and (A1) (dotted lines), respectively.

References

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, weibull and extreme value distributions to complex censored survival data using GLIM. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **29,** 156–163.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59,** 570–579.

Cai, T., Hyndman, R. J., and Wand, M. P. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics* **11,** 784–798.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society, Series B (Methodological)* **34,** 187–220.

Embury, S. H., Elias, L., Heller, P. H., Hood, C. E., Greenberg, P. L., and Schrier, S. L. (1977). Remission maintenance therapy in acute myelogenous leukemia. *Western Journal of Medicine* **126,** 267–272.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics* **36,** 299–305.

Miller, R. G. (1981). *Survival Analysis.* Wiley, New York.

PAPER IV

Approximate Bayesian model selection with the deviance statistic

Daniel Sabanés Bové & Leonhard Held

Approximate Bayesian model selection with the deviance statistic

Daniel Sabanés Bové* Leonhard Held†

Version: 16th February 2014

University of Zurich[UZH]

Institute of Social and Preventive Medicine
Division of Biostatistics

Bayesian model selection poses two main challenges: the specification of parameter priors for all models, and the computation of the resulting Bayes factors between models. There is now a large literature on automatic and objective parameter priors, which unburden the statistician from eliciting them manually in the absence of substantive prior information. One important class are g-priors, which were recently extended from linear to generalized linear models. To solve the computational challenge, we show that the resulting Bayes factors can conveniently and accurately be approximated by test-based Bayes factors (Johnson, 2008) using the deviance statistics of the models. For the estimation of the hyperparameter g, we show how empirical Bayes estimates correspond to shrinkage estimates from the

*E-mail: daniel.sabanesbove@ifspm.uzh.ch
†Corresponding author. E-mail: leonhard.held@ifspm.uzh.ch

literature, and propose a conjugate prior as a fully Bayes alternative. Connections to minimum Bayes factors are also discussed. We illustrate the methods with the development of a clinical prediction model for 30-day survival in the GUSTO-I trial, and with variable and function selection in Cox regression for the survival times of primary biliary cirrhosis patients.
Keywords: g-priors, shrinkage, variable selection, function selection, Bayes factor

1. Introduction

The problem of model and variable selection is pervasive in statistical practice. For example, it is central for the development of clinical prediction models (Steyerberg, 2009). For illustration, consider the famous GUSTO-I trial, which was a large randomised study for comparison of four different treatments in over 40 000 acute myocardial infarction patients (Lee et al., 1995). We will focus on a publicly available subgroup from the Western region of the USA with $n = 2188$ patients and prognosis of the binary endpoint 30-day survival (Steyerberg, 2009). In order to develop a clinical prediction model for this endpoint, we focus our analysis on the assessment of the effects of the covariates listed in Table 1. Among the 17 covariates, 4 are continuous (x_2, x_9, x_{10} and x_{16}), 2 are categorical (x_3 and x_{12}) and the remaining 11 are binary. We are interested both in identifying good predictors and obtaining reliable predictions. Since the response is binary, we will use logistic regression. We want to apply Bayesian inference to get posterior probabilities on variable inclusion and the most probable covariate effects.

There is now a large literature on automatic and objective Bayesian model selection, which unburden the statistician from eliciting manually the parameter priors for all models in the absence of substantive prior information (see *e. g.* Berger and Pericchi, 2001). This is also the situation we assume for the GUSTO-I data set. However, such objective Bayesian methodology is rather limited to the linear model (*e. g.* Bayarri, Berger, Forte, and García-Donato, 2012), due to computational and conceptual problems for non-Gaussian regression. One solution to this are test-based Bayes factors (Johnson, 2005), which we introduce now briefly. Consider a classical scenario with a

Variable	Description
y	Death within 30 days after acute myocardial infarction (Yes = 1, No = 0)
x_1	Gender (Female = 1, Male = 0)
x_2	Age [years]
x_3	Killip class (4 categories)
x_4	Diabetes (Yes = 1, No = 0)
x_5	Hypotension (Yes = 1, No = 0)
x_6	Tachycardia (Yes = 1, No = 0)
x_7	Anterior infarct location (Yes = 1, No = 0)
x_8	Previous myocardial infarction (Yes = 1, No = 0)
x_9	Height [cm]
x_{10}	Weight [kg]
x_{11}	Hypertension history (Yes = 1, No = 0)
x_{12}	Smoking (3 categories: Never / Ex / Current)
x_{13}	Hypercholesterolaemia (Yes = 1, No = 0)
x_{14}	Previous angina pectoris (Yes = 1, No = 0)
x_{15}	Family history of myocardial infarctions (Yes = 1, No = 0)
x_{16}	ST elevation on ECG: Number of leads (0–11)
x_{17}	Time to relief of chest pain more than 1 hour (Yes = 1, No = 0)

Table 1 *– Description of the variables in the GUSTO-I data set.*

null model nested within a more general alternative model. Traditionally, the use of Bayes factors requires the specification of proper prior distributions on all unknown model parameters of the alternative model, which are not shared by the null model. In contrast, Johnson (2005) defines Bayes factors using the distribution of a suitable test statistic under the null and alternative models, effectively replacing the data with the test statistic. This approach eliminates the necessity to define prior distributions on model parameters and leads to simple closed-form expressions for χ^2-, F-, t- and z-statistics. It can also be applied to nonparametric test statistics, see Yuan and Johnson (2008).

Classical usage of the value of a test statistic is for computing the corresponding P-value. Misinterpretation of P-values as posterior probabilities of the null hypothesis have led researchers to transform P-values to lower bounds on the corresponding Bayes factors (Edwards, Lindman, and Savage, 1963; Berger and Sellke, 1987; Goodman, 1999) and subsequently on the posterior probabilities of the null hypothesis. These methods are also based on Bayes factors using test statistics rather than the actual data and have thus much in common with the approach proposed by Johnson (2005). A slightly different route has been proposed in Sellke, Bayarri, and Berger (2001), who calibrate P-values by directly using the distribution of the P-value (rather than of a test statistic) under the two hypotheses.

The Johnson (2005) approach is extended in Johnson (2008) to the likelihood ratio test statistic and thus, if applied to generalized linear regression models (GLMs), to the deviance. This is explored further in Hu and Johnson (2009), where Markov chain Monte Carlo (MCMC) is used to develop a Bayesian variable selection algorithm for logistic regression. In this paper we build apon the work by Hu and Johnson (2009), combining g-prior methodology for the linear model with Bayesian model selection based on the deviance. This enables us to extend empirical (George and Foster, 2000) and fully Bayesian (Cui and George, 2008) approaches for estimating the hyperparameter g to GLMs. The approach provides a unified framework for objective Bayesian model selection and shrinkage of regression coefficients in GLMs and the Cox model. Note that estimation of regression coefficients is not discussed at all in the previous work on test-based Bayes factors. Links to the literature on calibration of the P-values using test statistics will also be explored.

The paper is structured as follows. In Section 2, we review the g-prior in the linear

and generalized linear model, and show that this prior choice is implicit in the application of test-based Bayes factors computed from the deviance statistic. In Section 3, we describe how the hyperparameter g influences the model and parameter inference, and introduce empirical and fully Bayes inference for it. Connections to the literature on minimum Bayes factors and shrinkage of regression coefficients are outlined. In Section 4, we discuss important issues for application of the methodology: variable and function selection, construction of an objective model prior and different ways to select or average the models. In Section 5, we apply the methodology in order to build a logistic regression model for predicting 30-day survival in the GUSTO-I trial, and compare our methodology with selected alternatives in a bootstrap study. Moreover, we select variables and functions in Cox regression for the survival times of primary biliary cirrhosis patients. In Section 6 we briefly discuss the extension of the approach to models with random effects.

2. Objective Bayesian model selection in regression

Consider a regression model $\mathcal{M}$ with linear predictor $\eta = \alpha + \boldsymbol{x}^\top \boldsymbol{\beta}$, from which we assume that the outcome $\boldsymbol{y} = (y_1, \ldots, y_n)$ was generated. We collect the intercept α, the regression coefficients vector $\boldsymbol{\beta}$ and possible additional (*e. g.* variance) parameters in $\boldsymbol{\theta}$. The models differ with respect to the content and the dimension of the covariate vectors $\boldsymbol{x}$, which we denote by the index $j \in \mathcal{J}$, where $\mathcal{J}$ is a finite index set. So a given model $\mathcal{M}_j$ defines a likelihood $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_j, \mathcal{M}_j)$.

With standard maximum likelihood estimation, we would estimate $\boldsymbol{\theta}_j$ by optimizing this likelihood and obtain the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}_j$. In Bayesian inference a prior distribution with density $\mathrm{p}(\boldsymbol{\theta}_j \mid \mathcal{M}_j)$ is assigned to the parameter vector $\boldsymbol{\theta}_j$. Its posterior density $\mathrm{p}(\boldsymbol{\theta}_j \mid \boldsymbol{y}, \mathcal{M}_j) \propto \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_j, \mathcal{M}_j)\, \mathrm{p}(\boldsymbol{\theta}_j \mid \mathcal{M}_j)$ is proportional to the product of the likelihood and the prior density, and maximizing this leads to the *maximum a posteriori* (MAP) estimate.

The importance of the parameter prior for model selection is immediately visible in the marginal likelihood

$$\mathrm{p}(\boldsymbol{y} \mid \mathcal{M}_j) = \int_{\boldsymbol{\theta}_j} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_j, \mathcal{M}_j)\, \mathrm{p}(\boldsymbol{\theta}_j \mid \mathcal{M}_j)\, d\boldsymbol{\theta}_j$$

of the model $\mathcal{M}_j$. This quantity is the first ingredient of the posterior model probabilities

$$\begin{aligned} \mathrm{p}(\mathcal{M}_j \,|\, \boldsymbol{y}) &= \frac{\mathrm{p}(\boldsymbol{y} \,|\, \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j)}{\sum_{j \in \mathcal{J}} \mathrm{p}(\boldsymbol{y} \,|\, \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j)} \\ &\propto \mathrm{p}(\boldsymbol{y} \,|\, \mathcal{M}_j)\,\mathrm{p}(\mathcal{M}_j) \\ &\propto \mathrm{BF}_{j,0}\,\mathrm{p}(\mathcal{M}_j), \end{aligned} \tag{1}$$

while the second ingredient is the prior model probability $\mathrm{p}(\mathcal{M}_j)$. Alternatively, the Bayes factor $\mathrm{BF}_{j,0} = \mathrm{p}(\boldsymbol{y} \,|\, \mathcal{M}_j) / \mathrm{p}(\boldsymbol{y} \,|\, \mathcal{M}_0)$ of model $\mathcal{M}_j$ *versus* a reference model $\mathcal{M}_0$ can be used. This ratio shows that improper priors may only be used for those parameters that are common to all models (*e. g.* here the intercept α), because only then the indeterminate normalising constant cancels in the posterior model probabilities.

In this paper, we propose to use a specific class of objective prior distributions $\mathrm{p}(\boldsymbol{\theta}_j \,|\, \mathcal{M}_j)$ for the model parameters. This prior is used for model selection problems, where it has proven advantages. The prior also induces shrinkage of $\boldsymbol{\beta}$, in the sense that the MAP estimate is a shrunken version of the MLE. Furthermore, it is an automatic prior, which means that it does not require specification of subjective prior information. It only depends on the model $\mathcal{M}_j$ under consideration. Note that also for the prior probabilities $\mathrm{p}(\mathcal{M}_j)$ on the model space an objective prior as that proposed in Section 4.2 can be used. We will now proceed to review the specific objective parameter prior family, namely the g-priors.

2.1. Zellner's g-priors and generalizations

We start with the original formulation of Zellner's g-prior for the Gaussian linear model in Section 2.1.1 and extend this to GLMs in Section 2.1.2.

2.1.1. Gaussian linear model

Consider the Gaussian linear model $\mathcal{M}_j$ with intercept α, regression coefficients vector $\boldsymbol{\beta}_j$ and variance σ^2, and collect all parameters in $\boldsymbol{\theta}_j = (\alpha, \boldsymbol{\beta}_j, \sigma^2)$. The likelihood obtained from n observations is

$$\mathrm{p}(\boldsymbol{y} \,|\, \boldsymbol{\theta}_j, \mathcal{M}_j) = \prod_{i=1}^{n} \mathrm{N}(y_i \,|\, \alpha + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j, \sigma^2),$$

where $\mathrm{N}(x \mid \mu, \sigma^2)$ denotes the univariate Gaussian density with mean μ and variance σ^2, and $\boldsymbol{x}_{ij} = (x_{i1}, \ldots, x_{id_j})^\top$ is the covariate vector for observation $i = 1, \ldots, n$. Using the $n \times d_j$ full rank design matrix $\boldsymbol{X}_j = (\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{nj})^\top$, we can rewrite this as a multivariate normal density

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}_j, \mathcal{M}_j) = \mathrm{N}_n(\boldsymbol{y} \mid \mathbf{1}_n \alpha + \boldsymbol{X}_j \boldsymbol{\beta}_j, \sigma^2 \mathbf{I}_n) \tag{2}$$

with $\mathbf{1}_n$ and $\mathbf{I}_n$ denoting the all-ones vector and identity matrix of dimension n, respectively. We assume that the covariates have been centered around 0, such that $\boldsymbol{X}_j^\top \mathbf{1}_n = \mathbf{0}_{d_j}$.

Zellner's g-prior (Zellner, 1986) specifies the Gaussian prior

$$\boldsymbol{\beta}_j \mid g, \sigma^2, \mathcal{M}_j \sim \mathrm{N}_{d_j}\left(\mathbf{0}_{d_j}, g\sigma^2 (\boldsymbol{X}_j^\top \boldsymbol{X}_j)^{-1}\right) \tag{3}$$

for the regression coefficients. For fixed σ^2, this can be interpreted as the posterior of the regression coefficients, if a locally uniform prior for $\boldsymbol{\beta}_j$ is combined with an imaginary sample $\boldsymbol{y}_0 = \mathbf{0}_n$ from the Gaussian linear model (2) with the same design matrix $\boldsymbol{X}_j$ but scaled residual variance $g\sigma^2$ rather than σ^2. This prior on $\boldsymbol{\beta}_j$ is usually combined with Jeffreys' prior on the intercept and variance parameters (Liang, Paulo, Molina, Clyde, and Berger, 2008):

$$\mathrm{p}(\alpha, \sigma^2) \propto \sigma^{-2}.$$

The posterior distribution of $(\alpha, \boldsymbol{\beta}_j^\top)^\top$ is then a multivariate t distribution, with posterior mean and mode for $\boldsymbol{\beta}_j$ given by

$$\mathbb{E}(\boldsymbol{\beta}_j \mid \boldsymbol{y}, g, \mathcal{M}_j) = \frac{g}{g+1}\hat{\boldsymbol{\beta}}_j = \frac{n \cdot \hat{\boldsymbol{\beta}}_j + n/g \cdot \mathbf{0}_{d_j}}{n + n/g} \tag{4}$$

This means that the MLE $\hat{\boldsymbol{\beta}}_j$, the ordinary least squares (OLS) estimate, is shrunk towards the prior mean zero. We call $t = g/(g+1)$ the shrinkage factor, which scales the MLE to obtain the posterior mean estimate (4). On the other hand, the posterior mean estimate is a weighted average of the MLE and the prior mean with weights proportional to the sample size n and the term n/g, respectively. Thus, n/g can be interpreted as the prior sample size, or $1/g$ as the relative prior sample size. The question how to choose or estimate g will be answered in Section 3.

One advantage of Zellner's g-prior is that the marginal likelihood, or equivalently the Bayes factor (BF) *versus* the null model $\mathcal{M}_0$, has a simple closed form expression in terms of the coefficient of determination R_j^2 of model $\mathcal{M}_j$ (Liang et al., 2008):

$$\mathrm{BF}_{j,0} = (1+g)^{(n-d_j-1)/2}\{1+g(1-R_j^2)\}^{-(n-1)/2}. \tag{5}$$

The closed form in terms of the statistic R_j^2 and the degrees of freedom d_j suggests that similar expressions can be derived for GLMs, a conjecture that will be confirmed in Section 2.2.

2.1.2. Generalized linear model

Now consider a GLM $\mathcal{M}_j$ with linear predictor $\eta_{ij} = \alpha + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j$, mean $\mu_{ij} = h(\eta_{ij})$ obtained with the response function $h(\eta)$, and variance function $v(\mu)$. The direct extension of the standard g-prior in the Gaussian linear model is then the generalized g-prior (Sabanés Bové and Held, 2011a)

$$\boldsymbol{\beta}_j \,|\, g, \mathcal{M}_j \sim \mathrm{N}_{d_j}\left(\mathbf{0}_{d_j}, g\, c\, (\boldsymbol{X}_j^\top \boldsymbol{W} \boldsymbol{X}_j)^{-1}\right), \tag{6}$$

where $\boldsymbol{W}$ is a diagonal matrix with weights for the observations (*e. g.* the binomial sample sizes for logistic regression). The constant $c = v\{h(0)\}h'(0)^{-2}$ preserves the interpretation of n/g as the prior sample size (*e. g.* $c = 4$ is obtained for logistic regression). It corresponds to the variance σ^2 in the standard g-prior (3) (Copas, 1983), which could also be formulated for general linear models, which have a non-unit weight matrix $\boldsymbol{W}$. As in Section 2.1.1, we specify Jeffreys' prior $\mathrm{p}(\alpha) \propto 1$ for the intercept α.

The connection between (6) and (3) is as follows. Denote the expected Fisher information (conditional on the variance σ^2 in the Gaussian linear model) for $(\alpha, \boldsymbol{\beta}_j^\top)^\top$ as $\mathcal{I}(\alpha, \boldsymbol{\beta}_j)$. In the Gaussian linear model, this $(d_j+1)\times(d_j+1)$ matrix is block-diagonal due to the centering of the covariates, and does not depend on the intercept nor the regression coefficients:

$$\mathcal{I}(\alpha, \boldsymbol{\beta}_j) = \begin{pmatrix} \mathcal{I}_{\alpha,\alpha} & \mathcal{I}_{\alpha,\boldsymbol{\beta}_j} \\ \mathcal{I}_{\alpha,\boldsymbol{\beta}_j}^\top & \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \end{pmatrix} = (\sigma^2)^{-1} \begin{pmatrix} n & \mathbf{0}_{d_j}^\top \\ \mathbf{0}_{d_j} & \boldsymbol{X}_j^\top \boldsymbol{X}_j \end{pmatrix}.$$

Now we see that (3) can also be written as

$$\boldsymbol{\beta}_j \,|\, g, \mathcal{M}_j \sim \mathrm{N}_{d_j}\left(\mathbf{0}_{d_j}, g \cdot \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}^{-1}\right). \tag{7}$$

In the GLM, $\mathcal{I}(\alpha, \boldsymbol{\beta}_j)$ depends on the parameters and is not necessarily block-diagonal. However, if we evaluate it at the prior mean $\alpha = 0$, $\boldsymbol{\beta}_j = \mathbf{0}_{d_j}$, it is indeed block-diagonal and $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} = c^{-1}\boldsymbol{X}_j^\top \boldsymbol{W}\boldsymbol{X}_j$. Therefore (6) can as well be written in the form (7).

In contrast to Gaussian linear models, the marginal likelihood resulting from use of the generalized g-prior does no longer have a closed form expression. For its computation, one has to resort to numerical approximations, *e.g.* a Laplace approximation. This requires a Gaussian approximation of the joint posterior $\mathrm{p}(\alpha, \boldsymbol{\beta}_j \mid \boldsymbol{y}, g, \mathcal{M}_j)$, which can be obtained with the Bayesian iteratively weighted least squares algorithm. See Sabanés Bové and Held (2011a, section 3.1) for more details.

2.2. Test-based Bayes factors

Based on the asymptotic behaviour of the deviance statistic in Section 2.2.1, we connect the resulting test-based Bayes factors with the g-prior in Section 2.2.2 and discuss the advantages over data-based Bayes factors in Section 2.2.3.

2.2.1. Asymptotic distributions of the deviance statistic

Consider the frequentist approach to model selection, where test statistics are used to assess the evidence against the null hypothesis $H_0 : \boldsymbol{\beta}_j = 0$ in a specific GLM $\mathcal{M}_j$. This null hypothesis restriction corresponds to the null model $\mathcal{M}_0$ without any covariates, so the linear predictor is identical to the intercept α. A popular choice is the deviance (or likelihood ratio test) statistic

$$z_j(\boldsymbol{y}) = 2 \log \left\{ \frac{\max_{\alpha,\boldsymbol{\beta}_j} \mathrm{p}(\boldsymbol{y} \mid \alpha, \boldsymbol{\beta}_j, \mathcal{M}_j)}{\max_{\alpha} \mathrm{p}(\boldsymbol{y} \mid \alpha, \mathcal{M}_0)} \right\}$$

Then we have the well-known result that in the case that the null hypothesis is true, *i.e.* conditional on $\mathcal{M}_0$, the distribution of the deviance $z_j(\boldsymbol{Y})$ converges for $n \to \infty$ to a chi-squared distribution $\chi^2(d_j)$ with d_j degrees of freedom.

What can be said about the asymptotic distribution of the deviance statistic under the alternative $H_1 : \boldsymbol{\beta}_j \neq 0$, *i.e.* in the unrestricted model $\mathcal{M}_j$? In order to answer this question, it has to be stated more precisely. We consider a sequence of local alternative hypotheses $H_1^n : \boldsymbol{\beta}_j = \mathcal{O}(1/\sqrt{n})$. That is, the absolute size of the true regression coefficients is scaled with $1/\sqrt{n}$, and thus gets smaller with increasing number of

observations n. This is the case of practical interest, because for larger $\boldsymbol{\beta}_j$, it would be trivial, and for smaller $\boldsymbol{\beta}_j$, it would be too difficult to differentiate between H_0 and H_1^n (Johnson, 2005, p. 691). In this setup, the distribution of the deviance converges for $n \to \infty$ to a non-central chi-squared distribution $\chi^2(d_j, \lambda_j)$ with d_j degrees of freedom, where $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j$ is the non-centrality parameter. Here $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$ denotes the expected Fisher information for $\boldsymbol{\beta}_j$ in model $\mathcal{M}_j$, evaluated at $\boldsymbol{\beta}_j = 0$. See Appendix A for a proof of this.

2.2.2. Defining the test-based Bayes factor

We now specify the generalized g-prior (7) for $\boldsymbol{\beta}_j$ in the alternative model $\mathcal{M}_j$. For the non-centrality parameter $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j$, this corresponds to the prior $\lambda_j \sim \mathrm{G}(d_j/2, 1/(2g))$ (also see Appendix A). From above we have the approximate "likelihood" $z_j \,|\, \lambda_j \overset{a}{\sim} \chi^2(d_j, \lambda_j)$ of the deviance statistic z_j. Johnson (2008, theorem 2) shows that the implied approximate marginal distribution of z_j is again a gamma distribution,

$$z_j \overset{a}{\sim} \mathrm{G}\left(\frac{d_j}{2}, \frac{1}{2(g+1)}\right).$$

which gives the the approximate "marginal likelihood" $\mathrm{p}_{\text{approx}}(z_j \,|\, \mathcal{M}_j)$ of model $\mathcal{M}_j$ in terms of the deviance statistic z_j. Furthermore, we have the approximate "marginal likelihood" $\mathrm{p}_{\text{approx}}(z_j \,|\, \mathcal{M}_0)$ of the null model $\mathcal{M}_0$ from $z_j \overset{a}{\sim} \mathrm{G}(d_j/2, 1/2)$. With these prerequisites, we can derive the test-based BF (TBF) (Johnson, 2008)

$$\mathrm{TBF}_{j,0} = \frac{\mathrm{p}_{\text{approx}}(z_j \,|\, \mathcal{M}_j)}{\mathrm{p}_{\text{approx}}(z_j \,|\, \mathcal{M}_0)} = (g+1)^{-d_j/2} \exp\left(\frac{g}{g+1}\frac{z_j}{2}\right). \tag{8}$$

of model $\mathcal{M}_j$ *versus* model $\mathcal{M}_0$. $\mathrm{TBF}_{j,0}$ approximates the data-based BF $\mathrm{BF}_{j,0} = \mathrm{p}(\boldsymbol{y} \,|\, \mathcal{M}_j)/\,\mathrm{p}(\boldsymbol{y} \,|\, \mathcal{M}_0)$ obtained with the same generalized g-prior (7). For example, in the Gaussian linear model we have $z_j = -n \log(1 - R_j^2)$ and obtain

$$\mathrm{TBF}_{j,0} = (g+1)^{-d_j/2}(1 - R_j^2)^{-tn/2}. \tag{9}$$

On the other hand, for large g we can approximate $1 + g(1 - R_j^2) \approx (1+g)(1 - R_j^2)$ in (5) and obtain

$$\mathrm{BF}_{j,0} \approx (1+g)^{-d_j/2}(1 - R_j^2)^{-(n-1)/2},$$

which is approximately equal to (9) for large g and n. A similar observation is made by Johnson (2008, section 3.2) for TBFs based on the F-statistic.

2.2.3. Advantages of the test-based Bayes factor

It is important to see that the TBF behaves like a data-based BF, in the sense that for a sequence of nested models $\mathcal{M}_0 \subset \mathcal{M}_1 \subset \mathcal{M}_2$, we have $\text{TBF}_{2,0} = \text{TBF}_{2,1} \cdot \text{TBF}_{1,0}$ (Hu and Johnson, 2009). Hence it is possible to compute posterior model probabilities from the TBF, by replacing the BF with the TBF in (1). These probabilities will be invariant to the choice of the baseline model $\mathcal{M}_0$, which we choose as the null model. This is also an advantage over a simple frequentist use of the deviance statistic: in the end we obtain posterior model probabilities, or other posterior probabilities of interest, *e. g.* inclusion probabilities. These are easier to interpret than the mere P-values in an analysis of deviance. Moreover, P-values are only suitable for pairwise model comparisons.

The TBF has several advantages over the data-based BF. First of all, it has a closed form in terms of the deviance statistic z_j, the prior variance factor g and the model dimension d_j. In contrast, the data-based BF needs to be approximated with numerical means, *e. g.* the Laplace approximation, and thus it does not have a closed form. The TBF formula (8) will allow us in Section 3 to study in detail the influence of g on shrinkage and model selection and to derive estimation procedures. Furthermore, the TBF can be computed more easily than the data-based BF, because it only requires the computation of the deviance statistic z_j. While this requires running an iteratively weighted least squares algorithm, it does not require the computation of the expected Fisher information $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} = c^{-1}\boldsymbol{X}_j^\top \boldsymbol{W}\boldsymbol{X}_j$, because that is just implicitly used in the prior formulation. In contrast, the Bayesian iteratively weighted least squares algorithm needs explicitly the inverse of the actual matrix $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$. Computation is even more preferable for the TBF if the goal is to estimate g.

Another advantage is that the TBF also works for the Cox proportional hazards model, where the nuisance parameter α now corresponds to an unspecified baseline hazard function. Banerjee (2005) shows that under the local alternative asymptotic framework from Section 2.2.1, the deviance (or partial likelihood ratio test) statistic in this special semiparametric model also follows a non-central chi-squared distribution with non-centrality parameter $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\boldsymbol{\beta}_j$, where $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$ is the efficient Fisher

information matrix evaluated at the null values. It was shown earlier (Murphy and van der Vaart, 2000) that under the null hypothesis, the deviance statistic follows approximately a central chi-squared distribution with d_j degrees of freedom. Since these asymptotic distributions have the same form as in the GLM case, the TBF formula can be used exactly in the same way for the Cox proportional hazards model. While the efficient Fisher information matrix, a generalization of the expected Fisher information to semiparametric models (Murphy and van der Vaart, 2000, p. 452), has a complicated analytic representation, this does not appear in the computations because it vanishes due to the assumption of the generalized g-prior (7). So far there has been no generalization of the g-prior to survival models, and we will see in Section 5.2 that the proposed TBFs give results which are close to those obtained from data-based BFs in a data-augmented Poisson model approximation of the Cox model.

3. Calibrating the prior

How does the prior variance factor g in the generalized g-prior (7) influence posterior inference? We will look at the implications on shrinkage and model selection in Section 3.1, and estimate g from the data using empirical Bayes (Section 3.2) and fully Bayes (Section 3.3) procedures.

3.1. The role of g for shrinkage and model selection

First, we look at the role of g for shrinkage. Assume first for simplification that the MLE $\hat{\boldsymbol{\theta}}_j = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_j^\top)^\top$ has asymptotically a $\mathrm{N}_{d_j+1}(\boldsymbol{\theta}_j, \mathcal{I}(0, \mathbf{0}_{d_j})^{-1})$ distribution, where $\boldsymbol{\theta}_j$ is the unknown true parameter and $\mathcal{I}(0, \mathbf{0}_{d_j})$ is the expected Fisher information matrix at the null values. This matrix is block-diagonal, *i.e.* $\mathcal{I}(0, \mathbf{0}_{d_j}) = \mathrm{diag}\{\mathcal{I}_{\alpha,\alpha}, \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\}$. Combining this Gaussian "likelihood" of $\boldsymbol{\theta}_j$ with the generalized g-prior, represented as partially improper Gaussian prior

$$\boldsymbol{\theta}_j \,|\, g, \mathcal{M}_j \sim \mathrm{N}_{d_j+1}\left(\begin{pmatrix} 0 \\ \mathbf{0}_{d_j} \end{pmatrix}, \begin{pmatrix} \infty & 0 \\ 0 & g \cdot \mathcal{I}^{-1}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \end{pmatrix}\right),$$

we obtain the approximate posterior distribution

$$\boldsymbol{\theta}_j \,|\, \boldsymbol{y}, g, \mathcal{M}_j \sim \mathrm{N}_{d_j+1}\left(\begin{pmatrix} \hat{\alpha} \\ t \cdot \hat{\boldsymbol{\beta}}_j \end{pmatrix}, \begin{pmatrix} \mathcal{I}^{-1}_{\alpha,\alpha} & 0 \\ 0 & t \cdot \mathcal{I}^{-1}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \end{pmatrix}\right). \tag{10}$$

Here $t = g/(g+1)$ is the same shrinkage factor for $\hat{\boldsymbol{\beta}}_j$ as in the Gaussian linear model from Section 2.1.1 (Copas, 1983).

The above assumption of the MLE distribution is mainly for deriving a simple form of the posterior distribution. In practice, we use in (10) the observed Fisher information matrix evaluated at the MLE instead of the expected Fisher information matrix evaluated at the null values. This should lead to a better accuracy of the approximation. In line with Copas (1983, section 8), we assume here that the observed Fisher information matrix is block-diagonal (see Section 3.3.3 for the implementation details). This is reasonable, because we have centered the covariate vectors in each model such that $\boldsymbol{X}_j^\top \mathbf{1}_n = \mathbf{0}_{d_j}$. For example, in logistic regression this corresponds to the assumption that not too many of the estimated probabilities $h(\hat{\eta}_{ij})$ are close to 0 or 1 (Copas, 1983, p. 327). This approach retains the assumption of approximate independence between the MLEs $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_j$.

While (10) holds exactly in the Gaussian linear model, it still holds approximately here in the non-Gaussian GLM. Thus we can again interpret g as the (approximate) ratio between the data sample size and the prior sample size. A smaller g leads to a smaller t and thus to stronger shrinkage of the $\boldsymbol{\beta}_j$ posterior to $\mathbf{0}_{d_j}$. In contrast, a larger value for g leads to t being closer to 1 and thus to weaker shrinkage. Note that also the approximate posterior covariance matrix for $\boldsymbol{\beta}_j$ is shrunk by the shrinkage factor t compared to the frequentist covariance matrix.

Shrinking the mean and covariance for the regression coefficients leads to "within-model-shrinkage" of the regression coefficients. By contrast, if Bayesian model averaging (BMA) of models differing with respect to included covariates is done, then there is a second form of shrinkage: If a covariate is not included in a model, then its coefficient is effectively estimated as zero in the full model. Taking into account these zeros leads to "between-model-shrinkage".

In order to understand the role of g for model selection, consider the TBF formula (8) and the limiting case of $g \to 0$. Then the generalized g-prior converges to a point mass at $\boldsymbol{\beta}_j = \mathbf{0}_{d_j}$, and thus $\mathcal{M}_j$ collapses to the null model $\mathcal{M}_0$. Consequently $\mathrm{TBF}_{j,0} \to 1$,

because both models are equal descriptions of the data in the limit. On the other extreme, the case $g \to \infty$ corresponds to an increasingly vague prior on $\boldsymbol{\beta}_j$. As is well known, arbitrarily inflating the prior variance of parameters that are not common to all models is not a safe strategy. Here we see immediately from (8) that $\mathrm{TBF}_{j,0} \to 0$ in this case. This means that no matter how well the model $\mathcal{M}_j$ fits the data compared to the null model $\mathcal{M}_0$, the latter is preferred if g is chosen large enough. This is an example of Lindley's paradox (Lindley, 1957).

In between these two extremes, quite a few fixed values for g have been recommended. The choice of $g = n$ corresponds to the unit information prior (Kass and Wasserman, 1995), where the relative prior sample size is $1/n$. Looking at the natural logarithm of the TBF, we find that for large n (Johnson, 2008, p. 358)

$$\begin{aligned}
\log(\mathrm{TBF}_{j,0}) &= -\frac{d_j}{2}\log(n+1) + \frac{n}{n+1}\frac{z_j}{2} \\
&\approx -\frac{d_j}{2}\log(n) + \frac{z_j}{2} \\
&= -\frac{1}{2}\{-z_j + d_j \log(n)\} \\
&= -\frac{1}{2}\mathrm{BIC}_j - \max_{\alpha} \log\{\mathrm{p}(y \mid \alpha, \mathcal{M}_0)\}
\end{aligned}$$

with $\mathrm{BIC}_j = -2\max_{\alpha,\boldsymbol{\beta}_j} \log\{\mathrm{p}(y \mid \alpha, \boldsymbol{\beta}_j, \mathcal{M}_j)\} + d_j \log(n)$. That means, the TBF is asymptotically ($n \to \infty$) equivalent to the Bayesian Information Criterion (BIC) and the corresponding approximation of the Bayes factor. Johnson (2008) and Hu and Johnson (2009) prefer a larger g, with $g/n \in [2, 6]$, and argue with favourable predictive properties and operating characteristics. Considering proposals for Zellner's g-prior in the Gaussian linear model, we mention the Risk Inflation Criterion (RIC) by Foster and George (1994), which corresponds to $g = d_j^2$, and the Benchmark prior by Fernández, Ley, and Steel (2001), which sets $g = \max\{n, d_j^2\}$. So the latter is equal to the unit information prior for $n > d_j^2$ and equal to the RIC otherwise.

3.2. Estimating g via empirical Bayes

Consider one specific model $\mathcal{M}_j$. If we choose g such that (8) is maximized, we obtain the estimate

$$\hat{g}_{\mathrm{LEB}} = \max\{z_j/d_j - 1, 0\}.$$

This is an empirical Bayes (EB) estimate because the prior parameter g is optimized in terms of the marginal likelihood $\mathrm{p}_{\mathrm{approx}}(z_j \,|\, \mathcal{M}_j)$. It is a local EB estimate because the estimate is specific for each model $\mathcal{M}_j$, $j \in \mathcal{J}$ (George and Foster, 2000). Using these values of g, the evidence in favour of the alternative hypotheses is maximized (Johnson, 2005, p. 693). This has the disadvantage that the resulting maximum TBFs

$$\mathrm{mTBF}_{j,0} = \max\left\{\left(\frac{z_j}{d_j}\right)^{-d_j/2} \exp\left(\frac{z_j - d_j}{2}\right), \quad 1\right\} \tag{11}$$

are not consistent if the null model is true (Johnson, 2008, p. 355), *i. e.* $\mathbb{P}(\mathcal{M}_0 \,|\, y) \not\to 1$ if $\mathcal{M}_0$ is true for $n \to \infty$. This is clear from above because (11) will always be larger than 1, instead of converging to 0, which is necessary for consistent accumulation of evidence in favour of the null model.

However, the corresponding shrinkage factors

$$\hat{t}_{\mathrm{LEB}} = \frac{\hat{g}_{\mathrm{LEB}}}{\hat{g}_{\mathrm{LEB}} + 1} = \max\{1 - d_j/z_j, 0\}. \tag{12}$$

are exactly the same as proposed by Copas (1997, p. 176), obtained with another rationale. He developed this formula specifically for logistic regression, by generalizing the formula for linear models. See also Van Houwelingen and Le Cessie (1990, p. 1322) for another justification of this shrinkage factor.

The maximum TBF has close connections to the Bayesian Local Information Criterion (BLIC) proposed by Hjort and Claeskens (2003, section 9.2). The only difference is that in the BLIC the deviance statistic is replaced by the squared Wald statistic for testing $\boldsymbol{\beta}_j = 0$. However, the squared Wald statistic shares the same non-central chi-squared distribution as the deviance statistic in the local asymptotic framework under the alternative model. Hence, the BLIC could be considered as a possibly even more computationally convenient approximation of the TBF in the sense of Lawless and Singhal (1978) who propose to replace the deviance statistic with the squared Wald statistic for model selection purposes. This comes at the price of losing the coherence of the TBF for nested models described in Section 2.2.3, which would hold only approximately.

There is also a close connection to minimum Bayes factors, which are used to transform P-values into a lower bound on the Bayes factor of the null *versus* the alternative model. As TBFs, these methods usually consider the value of a test statistic as the

data and transform this to (minimum) Bayes factors, thus quantifying the maximum evidence against a point null hypothesis. Key references are Edwards et al. (1963) and Berger and Sellke (1987) for normally distributed test statistics, see also Goodman (1999). A slightly different approach has been taken in Sellke et al. (2001), who directly use the distribution of the P-value (rather than considering the distribution of a test statistic generating the P-value) under the two hypotheses.

Of course, the deviance statistic z_j can also be transformed to a P-value $p_j = 1 - \mathrm{F}_{\chi^2(d_j)}(z_j)$ by applying the cumulative distribution function $\mathrm{F}_{\chi^2(d_j)}$ of the asymptotic chi-squared null distribution. As noted by Held (2010), depending on the degrees of freedom d_j, the maximum TBF (11) turns out to be equivalent to certain minimum Bayes factors:

1. For $d_j = 1$ it is equal to the Berger and Sellke (1987) bound for normal prior and test statistics.

2. For $d_j = 2$ it is equivalent to the Sellke et al. (2001) bound.

3. For $d_j \to \infty$ it is equal to the Edwards et al. (1963) universal bound for one-sided P-values obtained from normal test statistics. Moreover, it exactly equals the bound derived for a multivariate normal likelihood with known variance (Edwards et al., 1963, p. 234).

The proofs are given in Appendix B.

An alternative EB approach is to maximize the marginal likelihood over all models, *i. e.* to maximize

$$\mathrm{p}(\boldsymbol{y}) \propto \sum_{j \in \mathcal{J}} \mathrm{TBF}_{j,0}\, \mathrm{p}(\mathcal{M}_j) \tag{13}$$

with respect to g. The resulting estimate $\hat{g}_{\mathrm{GEB}}$ is the global EB estimate (Liang et al., 2008, section 2.4), which does not have a closed form expression and needs to be computed by numerical optimization of (13). It was investigated by George and Foster (2000) for the Gaussian linear model. From the computational side, calculating $\hat{g}_{\mathrm{GEB}}$ is more costly than calculating the model-specific $\hat{g}_{\mathrm{LEB}}$, and is even infeasible when $|\mathcal{J}|$ is very large. One solution could be to first perform a stochastic model search (see Section 4.3) and then restrict the sum in (13) to the set $\hat{\mathcal{J}}$ of models visited. The

stochastic model search could be based on the local EB estimates, say, and the resulting posterior model probabilities are then "corrected" by the global EB estimate.

The EB approach avoids arbitrary choices of g which may be at odds with the data. The local EB approach retains computational simplicity in comparison to the global EB approach. However, both ignore uncertainty about the estimates $\hat{g}_{\text{LEB}}$ and $\hat{g}_{\text{GEB}}$, respectively. In the next subsection we will perform fully Bayesian estimation of g and will thus be able to quantify the uncertainty about the estimate from its posterior distribution.

3.3. Full Bayes estimation of g

If we use a continuous hyperprior for g, then we obtain continuous mixtures of generalized g-priors, which we call generalized hyper-g priors (Sabanés Bové and Held, 2011a). Mixtures of g-priors for the Gaussian linear model were studied in detail by Liang et al. (2008). We use a hyperprior $\mathrm{p}(g)$ which is the same for all models, so the joint prior for the parameters and the models factorises as

$$\mathrm{p}(\alpha, \boldsymbol{\beta}_j, g, \mathcal{M}_j) = \mathrm{p}(\boldsymbol{\beta}_j \,|\, g, \mathcal{M}_j)\, \mathrm{p}(\mathcal{M}_j)\, \mathrm{p}(\alpha)\, \mathrm{p}(g).$$

3.3.1. A conjugate prior

In order to retain a closed form for the marginal likelihood of the model $\mathcal{M}_j$ when averaging (2.2.2) over the prior for g, the latter must be conjugate to the likelihood with kernel

$$\mathrm{p}_{\text{approx}}(z_j \,|\, g, \mathcal{M}_j) \propto (g+1)^{-d_j/2} \exp\left(-\frac{z_j/2}{g+1}\right).$$

From this we see that an inverse-gamma prior $\mathrm{IG}(a, b)$ on $g+1$, truncated appropriately to the range $(1, \infty)$, is conjugate (Cui and George, 2008, p. 891). The corresponding prior density function on g is

$$\mathrm{p}(g) = \mathrm{M}(a, b)(g+1)^{-(a+1)} \exp\left(-\frac{b}{g+1}\right), \tag{14}$$

where $\mathrm{M}(a, b) = b^a \{\int_0^b u^{a-1} \exp(-u)\, du\}^{-1}$ is the normalising constant. We denote this incomplete inverse-gamma distribution as $g \sim \mathrm{IncIG}(a, b)$. The model-specific

posterior density is

$$\begin{aligned} \mathrm{p}(g \mid z_j, \mathcal{M}_j) &\propto \mathrm{p}_{\text{approx}}(z_j \mid g, \mathcal{M}_j)\, \mathrm{p}(g) \\ &\propto (g+1)^{-(a+d_j/2+1)} \exp\left(-\frac{b+z_j/2}{g+1}\right), \end{aligned}$$

which shows that $g \mid z_j, \mathcal{M}_j \sim \text{IncIG}(a + d_j/2, b + z_j/2)$. Hence the marginal likelihood of model $\mathcal{M}_j$ is

$$\begin{aligned} \mathrm{p}(z_j \mid \mathcal{M}_j) &= \frac{\mathrm{p}_{\text{approx}}(z_j \mid g, \mathcal{M}_j)\, \mathrm{p}(g)}{\mathrm{p}(g \mid z_j, \mathcal{M}_j)} \\ &= \frac{\mathrm{M}(a,b) z_j^{d_j/2-1}}{\mathrm{M}(a+d_j/2, b+z_j/2) 2^{d_j/2} \Gamma(d_j/2)}, \end{aligned}$$

and dividing this with $\mathrm{p}_{\text{approx}}(z_j \mid \mathcal{M}_0)$ finally yields

$$\mathrm{TBF}_{j,0} = \frac{\mathrm{M}(a,b)}{\mathrm{M}(a+d_j/2, b+z_j/2)} \exp(z_j/2).$$

One analytic characteristic of the resulting model-specific posterior is the mode for the shrinkage factor t,

$$\mathrm{Mod}(t \mid z_j, \mathcal{M}_j) = \max\left\{1 - \frac{a+d_j/2-1}{b+z_j/2}, \quad 0\right\}. \tag{15}$$

This will be interesting when considering different choices of the hyperparameters a, b in Section 3.3.2.

If a non-conjugate prior on g is specified, the required integration of (2.2.2), $\mathrm{p}(z_j \mid \mathcal{M}_j) = \int \mathrm{p}_{\text{approx}}(z_j \mid g, \mathcal{M}_j)\, \mathrm{p}(g)\, dg$, can be performed by one-dimensional numerical integration. Two examples of non-conjugate hyperpriors on g which are used in the Gaussian linear model are the Zellner and Siow (1980) prior

$$g \sim \mathrm{IG}(1/2, n/2), \tag{16}$$

and the hyper-g/n prior proposed by Liang et al. (2008)

$$\frac{g/n}{g/n+1} \sim \mathrm{U}(0,1). \tag{17}$$

Both priors give considerable probability mass to g values proportional to n: The mode for the Zellner-Siow prior is $n/3$, and the median for the hyper-g/n prior is n.

3.3.2. Choosing the hyperparameters

The next question is then how to choose the hyperparameters a, b of the conjugate prior (14). Cui and George (2008) recommend $a = 1$ and $b = 0$, which leads to

$$t = \frac{g}{g+1} \sim \mathrm{U}(0,1), \tag{18}$$

i. e. a uniform prior on the shrinkage factor t. This is the hyper-g prior by Liang et al. (2008), a proper prior with normalising constant defined as the limit $\lim_{b\to 0} \mathrm{M}(a,b) = a$. The model-specific posterior mode (15) of t exactly equals the local EB estimate $\hat{t}_{\mathrm{LEB}}$ in (12). This is immediately clear because we have used a uniform prior on t, see (18). Moreover, the marginal posterior mode for t, taking into account all models, will equal the global EB estimate $\hat{t}_{\mathrm{GEB}}$. This indicates that using a hyper-g prior will lead to similar results as the EB methods.

Another choice of a, b can be motivated by the Zellner-Siow prior (16). We can approximate the Zellner-Siow prior by $g \sim \mathrm{IncIG}(a = 1/2, b = (n+3)/2)$, which has the same mode for g at $n/3$. We call this the Zellner-Siow adapted prior. For the model-specific posterior mode of t we obtain $\mathrm{Mod}(t \mid z_j, \mathcal{M}_j) = 1 - (d_j - 1)/(z_j + n + 3)$, which is always larger than $\hat{t}_{\mathrm{LEB}}$ in (12) and thus corresponds to weaker shrinkage of the regression coefficients.

The Zellner-Siow (adapted) prior depends on the sample size n, which leads to consistent model selection, even if the null model is true. Indeed, Johnson (2008) shows that for $g = \mathcal{O}(n)$ the TBF is consistent, because this prevents the alternative model from collapsing with the null model. Here we have prior mode $n/3$, which fulfils this condition. By contrast, the hyper-g prior (18) has its median at 1, which clearly does not fulfil the condition. Moreover, the model-specific posterior mode under the hyper-g prior equals the local EB estimate, which we have discussed in Section 3.2 to be inconsistent in case that the null model is true. The hyper-g/n prior (17) corrects this by scaling the prior to have median n. Therefore, if model or covariate identification is the primary goal of the analysis, either the Zellner-Siow adapted prior or the hyper-g/n prior should be used. However, these priors lead to weaker shrinkage than the local EB approach or the hyper-g prior. Stronger shrinkage is in general advantageous for prediction. Hence, if predicting new observations is the focus, local EB or the hyper-g prior are recommended.

For Cox proportional hazards models, Volinsky and Raftery (2000) note that there are only as many terms in the partial likelihood as there are uncensored observations, say n_{obs}. This is also the rate of growth of the efficient Fisher information matrix. Therefore Volinsky and Raftery (2000) propose a modified BIC which replaces the number of observations (n) with the number of uncensored observations (n_{obs}). This corresponds to another implicit prior on the parameters: instead of an "overall" unit information prior, it implicitly assumes an "uncensored" unit information prior. They go on to show that the revised BIC yields better results in an application. As the consistency results of Johnson (2008) rely on the fact that the prior covariance for $\boldsymbol{\beta}_j$ in the g-prior (7) stays asymptotically constant, we also use n_{obs} instead of n for the Zellner-Siow (adapted) priors and the hyper-g/n prior for the comparison of Cox proportional hazards models. Hence we obtain a hyper-g/n_{obs} prior. See Section 5.2 for an illustrating application.

3.3.3. Posterior parameter estimation

Given a model $\mathcal{M}_j$ with deviance statistic z_j, we would like to estimate the posterior distribution of its parameters $\boldsymbol{\theta}_j$. We do this by sampling from an approximation of the posterior distribution which avoids MCMC methods.

In order to take into account the uncertainty in the estimation of the hyperparameter g, we first sample from its marginal posterior $\mathrm{p}(g \mid z_j, \mathcal{M}_j)$. If a conjugate incomplete inverse-gamma prior distribution is specified for g, then the posterior is again of this form. Sampling from an $\text{IncIG}(a, b)$ distribution (14) is easy using inverse sampling via its quantile function

$$\mathrm{F}^{-1}_{\text{IncIG}(a,b)}(p) = \begin{cases} \frac{b}{\mathrm{F}^{-1}_{\text{IG}(a,1)}\{(1-p)\mathrm{F}_{\text{IG}(a,1)}(b)\}} - 1, & b > 0, \\ (1-p)^{-1/a} - 1, & b = 0. \end{cases}$$

which is given in terms of the quantile and cumulative distribution functions of the $\text{IG}(a, 1)$ distribution. If a non-conjugate prior is specified for g, then numerical methods can be used to sample from its approximate marginal posterior. Specifically we approximate the posterior density with a linear interpolation, which is a byproduct of the numerical integration to obtain the marginal likelihood of the model $\mathcal{M}_j$.

In the second step, we sample the actual model parameters $\boldsymbol{\theta}_j$ from their approximate posterior (10) given the sample for g. We use the observed Fisher information matrix, and delete the correlations between the intercept $\hat{\alpha}$ and the regression coefficients $\hat{\boldsymbol{\beta}}_j$ before inverting the matrix and scaling the lower right part with the sampled shrinkage factor $t = g/(g+1)$. The MLE $\hat{\boldsymbol{\beta}}_j$ is as well shrunk by t to obtain the mean of the conditional Gaussian distribution (10).

4. Implementation Issues

In this section we discuss some important implementation issues, which arise in the applications discussed in Section 5.

4.1. Variable and function selection

Model selection is often performed with two goals: variable and function selection. We will simplify the exposition by considering only Gaussian examples in this Section.

In the Bayesian framework, variable selection is the task of assessing the importance of a set of independent variables (covariates) by computing posterior probabilities of their inclusion in the regression model. Starting with p covariates x_k, $k = 1, \ldots, p$, we consider different models $\mathcal{M}_j$ with means $\mu_{ij} = \alpha + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j$. The models differ in the definition of the design vectors $\boldsymbol{x}_{ij}$, which are subvectors of the full design vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ comprising all p covariates. Define the binary indicator γ_{jk}, which is 1 if the covariate x_k is contained in the model $\mathcal{M}_j$ and 0 otherwise. Then based on all posterior model probabilities $\mathrm{p}(\mathcal{M}_j \mid \boldsymbol{y})$, the posterior variable inclusion probabilities

$$\mathbb{P}(x_k \in \mathcal{M} \mid \boldsymbol{y}) = \sum_{j \in \mathcal{J}} \gamma_{jk} \, \mathrm{p}(\mathcal{M}_j \mid \boldsymbol{y})$$

can be computed.

However, if there are continuous variables, just selecting among them is often not enough: The question is how to choose the functional forms of their effects. To ease notation, let all covariates be continuous, then we consider the additive model $\mu_{ij} = \alpha + \sum_{k=1}^{p} \gamma_{jk} f_{jk}(x_{ik})$. This function selection can be seen as a second hierarchy level in the model selection problem: Given a set of covariates, *i. e.* conditional on the variable selection γ_{jk}, how to select the functions f_{jk}? Of course we would like

to allow linear effects $f_{jk}(x_{ik}) = x_{ik}\beta_{jk}$, but we would also like to compare this with models having non-linear effects $f_{jk}(x_{ik})$. One simple class of non-linear parametric covariate transformations are the fractional polynomials (FPs, Royston and Altman, 1994). Bayesian inference for the resulting normal linear models was developed by Sabanés Bové and Held (2011b), and extended to GLMs by Sabanés Bové and Held (2011a). The idea of the FPs is to improve upon the manual ad-hoc inclusion of *e.g.* quadratic covariate terms in the model by automatic inclusion of not only quadratic, but also cubic, square root, logarithm, and reciprocal terms. Nevertheless, with appropriately chosen design covariates all models reduce to linear models, such that computations are not more difficult than for plain variable selection. An alternative to FPs are splines, which have been used with generalized g-priors by Sabanés Bové, Held, and Kauermann (2012). However, they require random effects for the spline coefficients, which excludes the use of TBFs, see Section 6 for more discussion of this issue. Moreover, for Cox models a Poisson approximation (Sabanés Bové and Held, 2013) has to be used, which greatly increases the computational costs.

4.2. Model prior

In the absence of subjective prior information on the importance of covariates, we use the following model prior. Assume that the binary variable inclusion indicators γ_{jk} are independent and identically Bernoulli distributed with probability π for inclusion. If we denote the number of covariates included in model $\mathcal{M}_j$ as $d_j = \sum_{k=1}^{p} \gamma_{jk}$, we have $d_j \sim \text{Bin}(p, \pi)$. Assigning a uniform prior $\pi \sim \text{U}(0,1)$ to the inclusion probability leads to a marginal uniform prior on d_j, which is an intuitively sensible prior assumption (Geisser, 1984). However, the separate indicators γ_{jk} are dependent after integrating out π. Therefore this prior is multiplicity-corrected (Scott and Berger, 2010).

For additional function selection, we extend it as follows. Conditional on the selection of d_j variables, the number of non-linearly modelled covariate effects is uniformly distributed on $\{0, 1, \ldots, d_j\}$. The selection of the corresponding covariates and the choice of non-linear functions is then uniformly distributed on all possible configurations and transformations, respectively. This prior leads to a marginal prior inclusion probability of 50% for each covariate, with 25% each for a linear and a non-linear

effect.

4.3. Model selection for prediction

Exact computation of the posterior probability of any model $\mathcal{M}_j$ via (1) requires the computation of all unnormalised model probabilities $\mathrm{BF}_{j,0}\,\mathrm{p}(\mathcal{M}_j)$ in the model space $\mathcal{J}$. This "exhaustive" evaluation of the model space is only possible for applications with a relatively small number p of covariates, because the number of models in $\mathcal{J}$ grows with 2^p for variable selection, and even faster for combined variable and function selection. As an alternative, stochastic search algorithms have been developed. Such algorithms are basically MCMC samplers on the model space. Instead of waiting until convergence of the Markov chain, one uses a relatively small number of samples of models to constitute a set $\hat{\mathcal{J}}$ of promising models, which is plugged in for the original model space $\mathcal{J}$. Normalisation of model probabilities, computation of posterior variable inclusion probabilities, *etc.* is then based on $\hat{\mathcal{J}}$ instead of $\mathcal{J}$. We use the stochastic search algorithm described in Sabanés Bové and Held (2011b) for variable selection and FP models, and the algorithm in Sabanés Bové et al. (2012, section 4) for spline models.

When aiming for good predictions, the question is how to proceed from the full set of models $\mathcal{J}$ or a promising subset $\hat{\mathcal{J}}$. If a single model is required, then the MAP model, *i. e.* the model with the highest posterior model probability, is the traditional choice. However, there are alternatives which take more than just one model into account. The median probability model (MPM) was defined by Barbieri and Berger (2004) for the variable selection problem as the model comprising exactly those covariates which have inclusion probabilities greater than or equal to 50%. The MPM has attractive theoretical properties if the covariates can be assumed to come from some distribution, and minimizing the expected predictive squared error loss also with respect to this covariate distribution is the goal. When the covariates can instead be considered fixed, BMA should be used, because it minimizes the predictive squared error loss (Barbieri and Berger, 2004). To make BMA feasible, often the set of models over which to average is reduced, and in the applications we take a fixed number of best models. For combined variable and function selection, the concepts of MAP model and MPM can be extended by grouping together models that only differ in the

functional form of included covariate effects. The resulting groups, or meta-models, can then produce predictions by averaging over the individual sub-model predictions, weighted appropriately with the posterior model probabilities. We call these methods MAP-BMA and MPM-BMA, respectively.

5. Applications

We illustrate the performance of the test-based Bayes factors in comparison to the data-based Bayes factors with logistic regression in Section 5.1 and with Cox regression in Section 5.2. Note that the TBF methodology is implemented in the efficient R-package "glmBfp" available from R-Forge.[1]

5.1. Prognostic modelling of 30-day survival

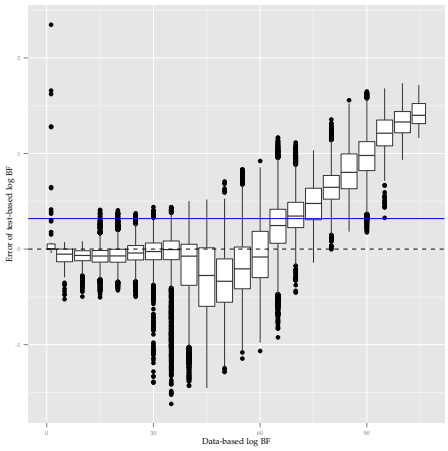
We would now like to develop a prediction model for 30-day survival for the GUSTO-I trial data introduced in Section 1.

5.1.1. Variable selection

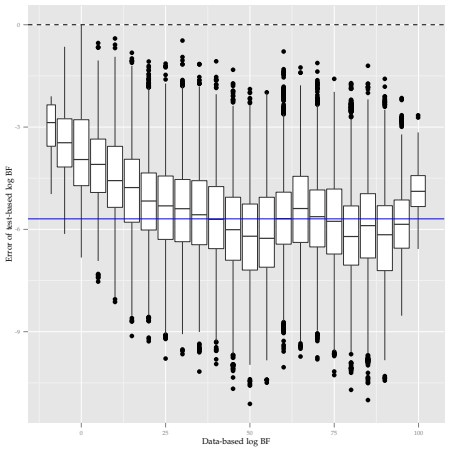
We start with variable selection. As there are $p = 17$ covariates in the data set, there are $|\mathcal{J}| = 2^{17} = 131\,072$ different models. Since this is still a manageable size, we can evaluate all models. We will use this example to demonstrate the good approximation of the data-based analysis by the test-based approach, where the accuracy differs between the different estimation methods for g.

In Figure 1 we compare the resulting log BFs between the test-based and the data-based approaches, when local EB, the Zellner-Siow adapted prior, the hyper-g prior or the hyper-g/n prior are used. We see that the results are quite close, although there are larger errors of the test-based Bayes factors for the Zellner-Siow adapted prior (Figure 1b). Note that we also use the Zellner-Siow adapted prior for the data-based BFs. While for the other prior choices, we observe a relative increase of the test-based log BFs for higher data-based log BFs, it is interesting that we see the other tendency for the Zellner-Siow adapted prior.

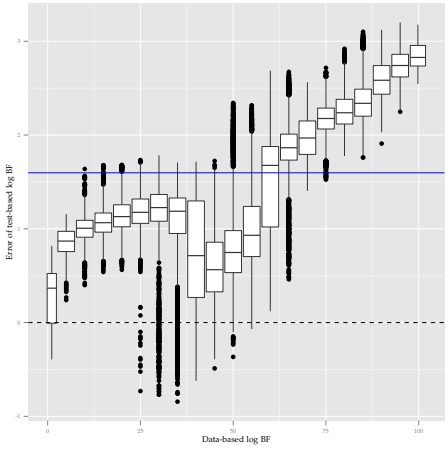
[1] To install the R-package, just type `install.packages("glmBfp",repos="http://r-forge.r-project.org")` into R.



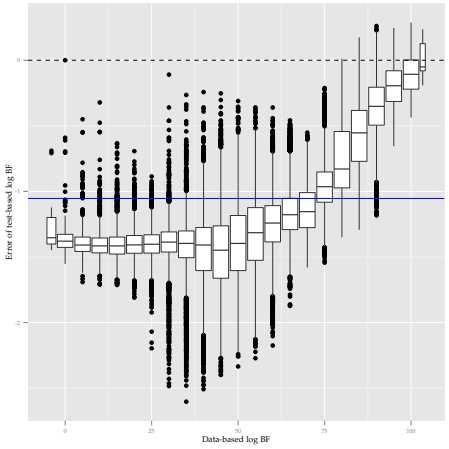
(a) Comparison of log BFs with local EB (mean difference: 0.319).



(b) Comparison of log BFs with Zellner-Siow adapted prior (mean difference: −5.691).

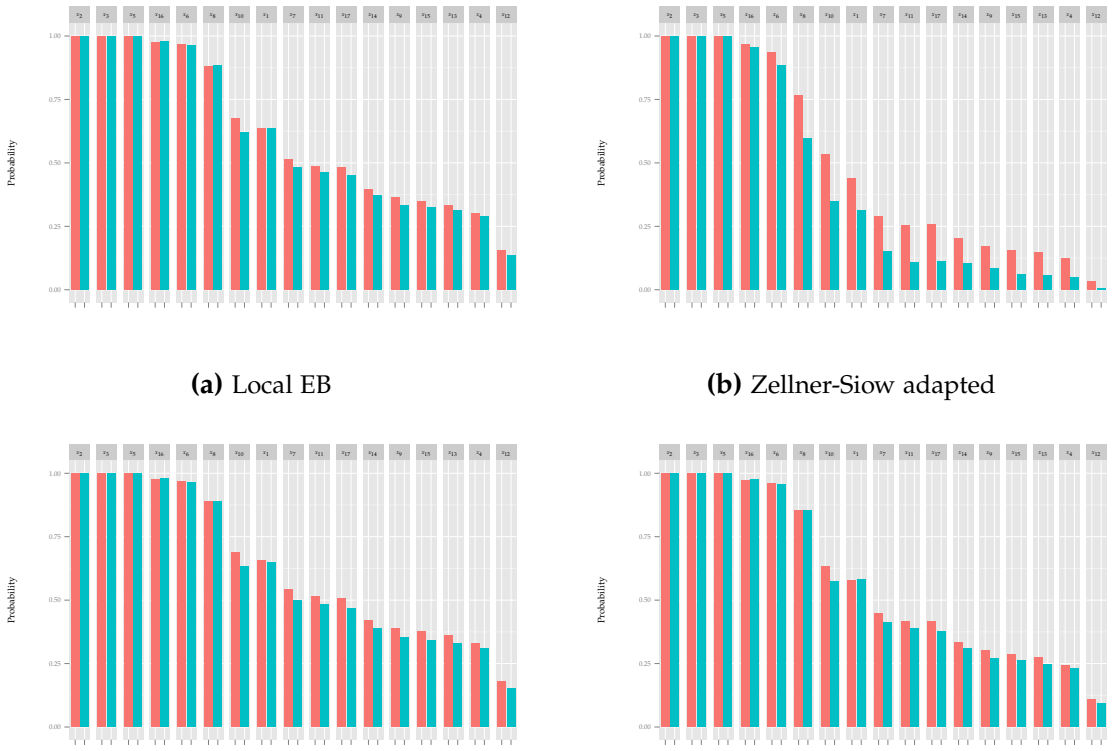


(c) Comparison of log BFs with hyper-g prior (mean difference: 1.597).



(d) Comparison of log BFs with hyper-g/n prior (mean difference: −1.054).

Figure 1 – *GUSTO-I data: Comparing test-based and data-based log BFs. The means of the errors of the test-based log BF are given in the captions (solid blue line).*



(a) Local EB

(b) Zellner-Siow adapted

(c) Hyper-g

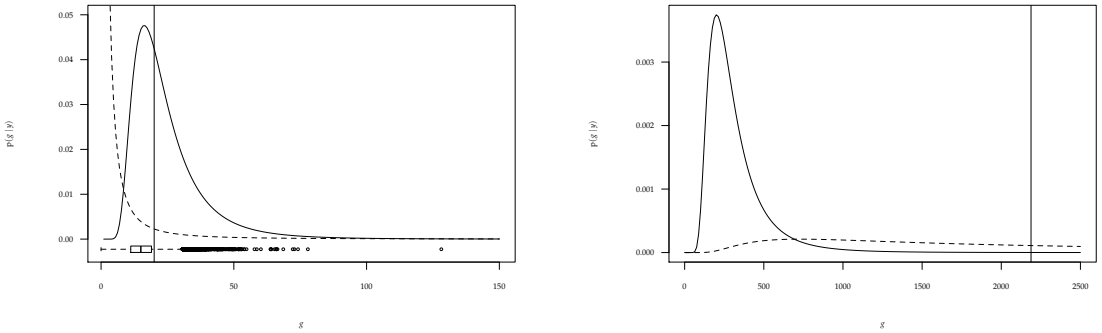
(d) Hyper-g/n

Figure 2 – *GUSTO-I data: Inclusion probabilities for all approaches, comparing the fully Bayes (left bars, ■) and the TBF approach (right bars, ■). The covariates are ordered with respect to the results from the data-based Bayes factors for the hyper-g/n prior.*

Following the similarities of the Bayes factors, it is not astonishing that the resulting posterior variable inclusion probabilities (under the multiplicity-corrected model prior from Section 4.2) are very similar between the test-based and data-based analyses, see Figure 2. The two neighbouring bars have almost the same height for all covariates and for all settings except the Zellner-Siow adapted prior, where the differences are larger. However, there are substantial differences between the four different settings for estimating g.

The computations were between 11 (local EB) and 50 (Zellner-Siow adapted) times faster for the test-based BFs compared to the data-based BFs. This illustrates one main advantage of the TBFs, namely computational efficiency.

In Figure 3 the posterior distributions on g are compared with the underlying conjugate prior distributions (Zellner-Siow adapted and hyper-g) and local as well as global EB estimates of g. We clearly see the difference between the two priors resulting from the different hyperparameter choices. The BIC choice $g = n$ is not supported by the data, as all estimates are far below this value. Local EB estimates give the smallest values for g, together with the posterior mode of g under the hyper-g prior and the global EB estimate. The posterior mode of g under the Zellner-Siow adapted prior is larger.



(a) Hyper-g prior and resulting posterior, together with local EB (boxplot for the values at bottom of the plot) and global EB (vertical line) estimates of g.

(b) Zellner-Siow adapted prior and resulting posterior, together with $g = n$ (vertical line).

Figure 3 – *GUSTO-I data: Comparison of posteriors (solid lines) and priors (dashed lines) from the conjugate incomplete inverse-gamma prior on g with hyper-g (left) and Zellner-Siow adapted (right) hyperparameter choices.*

5.1.2. Variable and function selection

Since 4 covariates are continuous, we now also want to consider the possibility of non-linear covariate effects, using FPs (see Section 4.1). We choose the hyper-g/n prior (17) because we are mainly interested in model selection here.

Due to the huge size of the model space ($|\mathcal{J}| = 2^{13} \cdot 45^4 > 3 \cdot 10^{10}$), it is no longer possible to evaluate all models. We therefore used stochastic model search with 100 000 iterations to traverse the model space and find good models. Using data-based Bayes

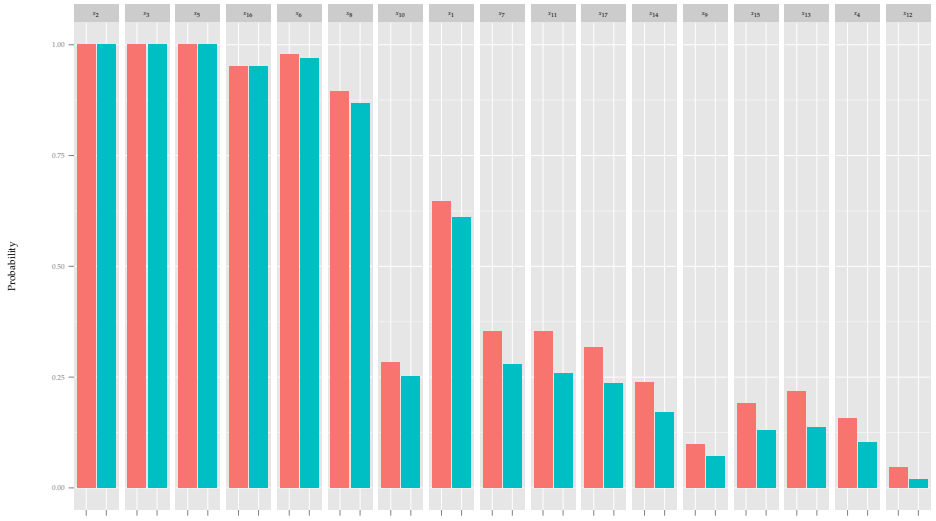


Figure 4 – *GUSTO-I data: Inclusion probabilities from the FP models, comparing the data-based (left bars, ■) and the test-based Bayes factor approaches (right bars, ■). The covariates are ordered as in Figure 2.*

factors, this took 513 minutes. The resulting MAP model includes the discrete covariates x_1, x_3, x_5, x_6 and x_8 and the continuous FPs x_2^3 and x_{16}^1, and has posterior model probability $3.90 \cdot 10^{-3}$. When we instead use test-based Bayes factors, it took only 11 minutes and we obtained the MAP model with the discrete covariates x_1, x_3, x_5, x_6 and x_8 and the continuous FPs x_2^1 and x_{16}^1, with posterior model probability $8.69 \cdot 10^{-3}$. The slight difference in the MAP model configurations is not worrying, because the two models are on places 1 and 3 in both approaches, the second best model being identical between data-based and test-based results. While the ranking is hence similar, the posterior model probabilities are larger in the test-based results ($8.50 \cdot 10^{-3}$ and $6.73 \cdot 10^{-3}$ for the second and third best model) than in the data-based results ($3.78 \cdot 10^{-3}$ and $2.99 \cdot 10^{-3}$, respectively).Also Steyerberg (2009, p. 103) investigated the age covariate x_2 and found that there is a "reasonably linear relationship" with the outcome, which supports the simpler FP x_2^1 in the MAP model of the TBF results.

We compare the posterior inclusion probabilities between both approaches in Figure 4. Overall, the inclusion probabilities are quite similar between the two methods, with the test-based ones always smaller than the data-based ones. Comparing the in-

clusion probabilities with those from Figure 2, we see that x_{10} (the weight variable) has a drastically smaller inclusion probability when FP transformations are considered. It falls out of the median probability model with now only about 25% inclusion probability. Also the inclusion probability of x_9 (the height variable) is relatively smaller here. This might be due to the used model flexibility for the age variable x_2, which has a probability of 91.1% for non-linear inclusion in the top 3000 models saved from the stochastic model search. The previously seen effects of x_9 and x_{10} may be surrogates for the non-linear effect of x_2.

5.1.3. Bootstrap study

For quantifying the predictive performance of our variable and function selection methods, we used the area under the ROC curve (AUC, measures discrimination), the logarithmic scoring rule (LS) and the Brier scoring rule (BS) (both are measuring discrimination and calibration). The apparent performance of the methods, *i. e.* when fitting the original sample and then predicting it, is well-known to be of little value for estimating the prediction performance for new data. Therefore we estimated the out-of-sample discrimination and calibration of the methods using optimism-corrected bootstrap estimates (Steyerberg, Harrell, Borsboom, Eijkemans, Vergouwe, and Habbema, 2001) of AUC, LS and BS: For every bootstrap iteration, we fitted the methods to the bootstrap sample, and then predicted both the original sample and the bootstrap sample. The average difference between the performances in the bootstrap sample and the original sample is an estimate of the optimism in the apparent performance. The final estimate of the out-of-sample performance is then defined as the apparent performance minus the optimism estimate.

Since the focus is now more on obtaining good predictions rather than identifying influential covariates, we use the hyper-g prior (18). This leads to stronger shrinkage than the hyper-g/n prior, and hence to better out-of-sample predictions. Moreover, we use a BMA over the best 3000 models found under the two approaches, and also apply the MAP-BMA and MPM-BMA versions described in Section 4.3. Moreover, we want to compare our methods with generalized linear (GLM) and additive models (GAM) (Wood, 2011), for which a variable selection method was proposed by Marra and Wood (2011, section 2.1). A non-Bayes approach for selecting FPs (MFP) is de-

scribed by Sauerbrei, Royston, and Binder (2007). Ours and these alternative methods can either be used as plain variable selection, or for combined variable and function selection. We also fitted a GAM version of the Lee et al. (1995) model, adapted to the available covariates in our data set, including linear or dummy-coded effects for $x_3, x_4, x_5, x_6, x_7, x_8, x_{11}$ and x_{12}, a smooth term for x_9 and a factor-smooth interaction for x_2 and x_3. The results are shown in Table 2.

Comparing first the performances within the stochastic searches, we see that almost always BMA is better than MPM, and MPM is better than MAP, the only exception being combined variable and function selection with data-based BFs where MAP and MAP-BMA are better than BMA and MPM-BMA. This is not surprising, given the theoretical advantage of BMA over single models concerning prediction.

Second, the test-based methods perform similarly well as the data-based ones for plain variable selection. For combined variable and function selection, the test-based methods perform worse, with the distance being largest for the MAP choice. This is reasonable as the errors of the TBFs have more impact when a single model is selected, based only on its TBF. On the other hand, for MPM-BMA and BMA, the test-based results come closer the data-based ones. It is intuitively clear that the data-based methods yield better predictions for the new data, because they do not suffer from the additional approximation due to the use of the deviance statistic.

Third, the methods based on the generalized hyper-g prior and BMA are better than the alternative methods. Among the alternative methods, the GLM with selection performs best for the variable selection exercise. Its additional flexibility from separate shrinkage of the coefficients leads to a better performance than the MAP models for. However, for combined variable and function selection, the data-based methods are better than GAM with selection. Here, the full GLM/GAM and the MFP selection methods yield worse results.

Fourth, for the data-based BFs, using combined variable and function selection yields a clear advantage in terms of better predictions in this example, when we compare the results with and without function selection. However, for the test-based BFs and the alternative methods, this does not hold. This in accordance to the experience of Ennis, Hinton, Naylor, Revow, and Tibshirani (1998), who compared different non-linear methods with the original Lee et al. (1995) model and found that the latter could not be outperformed. In our study, the adapted Lee model performed badly, which

might be explained by the fact that we could not use the original Lee model due to missing covariates in our data set.

In supplementary material we also give analogous tables with the results obtained with local EB estimation of g and with the hyper-g/n prior.

While the hyper-g/n results are similar to Table 2, we observe a decreased performance of the data-based BFs with the local EB approach. In summary, the results indicate that for selecting a single model (either MAP or MPM) with test-based BFs, then the hyper-g/n prior yields better predictions than hyper-g and local EB. This might be explained by the model selection consistency property of the hyper-g/n prior. For BMA, the differences are smaller, with local EB having slightly better results. Overall, we observed the best predictions in the data-based MAP-BMA with combined variable and function selection and local EB estimation of g (AUC 0.8541, LS 392.1 and BS 105.9).

5.2. Primary Biliary Cirrhosis

To illustrate the applicability of the proposed methodology in Cox regression, we consider survival data provided by Therneau and Grambsch (2000) on primary biliary cirrhosis (PBC) patients, from which we use the $n = 276$ complete observations. There are $n_{\text{obs}} = 111$ survival times which have been observed without censoring. Table 3 contains a description of the variables in the data set.

5.2.1. Variable selection

In Figure 5 we compare posterior variable inclusion probabilities based on the hyper-g/n_{obs} prior. If we used the total number of observations n instead of n_{obs}, then the inclusion probabilities are lower. The reason is that the prior favours larger values for g, which corresponds to smaller Bayes factors of the alternative models *versus* the null model (compare Section 3.1). For the model space, we use the multiplicity-corrected model prior from Section 4.2.

We compare results based on TBFs (plain variable selection and combined with FP function selection) with results obtained from data-based BFs with a Poisson approximation (Sabanés Bové and Held, 2013) (plain variable selection and combined with splines modelling of covariate effects).

			AUC	LS	BS
Variable selection	Data-based	MAP	0.8404	398.5	107.6
		BMA	0.8449	394.3	106.8
		MPM	0.8435	394.8	106.8
	Test-based	MAP	0.8404	399.2	107.7
		BMA	0.8448	394.5	106.6
		MPM	0.8425	397.7	107.4
	Alternatives	GLM (select)	0.8447	395.4	107.0
		GLM (full)	0.8445	395.7	106.9
		MFP (linear)	0.8413	399.1	107.6
		Adapted Lee	0.8369	400.3	108.0
Transformations	Data-based	MAP	0.8446	392.6	105.9
		BMA	0.8432	393.1	106.1
		MAP-BMA	0.8449	392.1	105.9
		MPM-BMA	0.8417	394.9	106.7
	Test-based	MAP	0.8364	402.6	108.3
		BMA	0.8420	395.0	106.6
		MAP-BMA	0.8398	397.5	107.2
		MPM-BMA	0.8399	397.4	107.2
	Alternatives	GAM (select)	0.8406	397.9	107.5
		GAM (full)	0.8376	401.2	107.9
		MFP	0.8388	401.7	108.4

Table 2 – *GUSTO-I data: Comparison of the predictive performance of plain variable selection (first half) and combined variable and FP function selection (second half) methods with alternative approaches, using bootstrap estimates (300 samples) of AUC, Log-score (LS) and Brier score (BS). The LS is defined as* $-\sum_{i=1}^{n} \log\{\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}\}$*, where* $\hat{\pi}_i$ *is the predicted probability of death for the ith patient. The BS is defined as* $\sum_{i=1}^{n}(y_i-\hat{\pi}_i)^2$*. For the bootstrap runs, only* 30 000 *iterations of the stochastic search were done.*

Variable	Description
y	Number of days between registration and death / transplantation / end of study
δ	0 for censored (including transplantation), 1 for observed survival time
x_1	Treatment? (1: D-penicillamin, 2: Placebo)
x_2	Age [years]
x_3	Gender? (0: Male, 1: Female)
x_4	Presence of ascites? (0: No, 1: Yes)
x_5	Presence of hepatomegaly or enlarged liver? (0: No, 1: Yes)
x_6	Blood vessel malformations in the skin? (0: No, 1: Yes)
x_7	Edema? (0: no, 0.5: not or successfully treated, 1: therapy-resistant)
x_8	Serum bilirubin [mg/dl]
x_9	Serum cholesterol [mg/dl]
x_{10}	Serum albumin [g/dl]
x_{11}	Urine copper [ug/day]
x_{12}	Alkaline phosphotase [U/l]
x_{13}	Aspartate aminotransferase, once called SGOT [U/ml]
x_{14}	Triglycerides [mg/dl]
x_{15}	Platelet count
x_{16}	Standardised blood clotting time
x_{17}	Histologic stage of disease? (1, 2, 3, 4)

Table 3 *– Description of the variables in the Primary Biliary Cirrhosis data set.*

It is interesting that the results of the two plain variable selection methods shown in Figure 5a are very similar. We see that the covariates x_2, x_7, x_8, x_{10}, x_{11}, x_{13}, x_{16} and x_{17} have notably higher values than the other covariates. A similar selection is produced *e. g.* by the adaptive LASSO (Zhang and Lu, 2007). Note the drastic computational performance difference between the two methods: While we could compute 291 TBFs per second, we could only compute about 11 data-based BFs per minute. This means that the Cox approach with TBFs is around 1500 times faster than the Poisson approximation with data-based BFs. This is due to the huge blow-up of the pseudo Poisson data set (pseudo sample size is 37 155) required for the latter approach compared to the original data set ($n = 276$); see Sabanés Bové and Held (2013).

5.2.2. Variable and function selection

When FP transformations of the covariates are considered, the overall picture changes (Figure 5b): x_{11}, x_{13} and x_{17} have now small posterior inclusion probabilities, while x_2, x_7, x_8, x_{10}, and x_{16} keep their high inclusion probabilities. Using either the Cox partial likelihood or the Poisson approximation with TBF gives almost the same results here. Again, note the difference in computation times: While with the Cox approach we could compute 318 models per second, this number is only 4 for the Poisson likelihood approximation. Even slower is computing the data-based BFs for the Poisson approximation, then we could only compute 8 models per minute. The results of the latter method are close to the one with Poisson and TBF, except for x_7 where the inclusion probability is clearly lower.

Finally, we would like to compare the results obtained with the parametric FP transformations with the results from spline modelling of the covariate effects. Figure 5c shows the posterior inclusion probabilities obtained with both approaches. The probabilities are similar, except for larger discrepancies for x_7, x_{11}, x_{16} and x_{17}: While x_{16} has inclusion probability of more than 50% using the FPs, it has less than 10% using splines. On the other hand, x_{17} has inclusion probability below 10% using FPs, while it has 40% using splines. Note that the spline-based approach is computationally very expensive. It could on average only process 4 models per minute in this example.

In Figure 6 we compare the estimated covariate effects in the MAP models. For x_8 (Figure 6b), both methods fit a very similar non-linear effect, although there appears

to be an artefact of the FP (including the terms x_8^{-1} and x_8^{-2}) for values close to zero. For x_{10} (Figure 6c), both methods fit a very similar linear effect. For x_2 (Figure 6a), the FP method selected a linear effect, while the non-linear effect fitted by the splines method is more pronounced for lower values and reaches a plateau around 65 years. In the splines model, having edema despite diuretic therapy (from x_7, using a dummy covariate) gives a higher estimated (log) hazard, with a posterior expected increment of 1.08 on the log hazard (95% credible interval: 0.4 to 1.68). In the FP model, instead of x_7, the standardised blood clotting time x_{16} is positively associated with a higher risk of death (Figure 6d).

6. Discussion

In this paper we showed how test-based Bayes factors approximate data-based Bayes factors. Specifically, we looked at Bayes factors derived from the deviance statistic for generalized linear and Cox models, and exposed the fact that the implicitly used prior on the regression coefficients is a generalized g-prior. As with the data-based Bayes factors, estimation of g is possible and recommended. Local EB estimation of g leads to posterior means of the regression coefficients that correspond to shrinkage estimates from the literature. Alternatively, full Bayes treatment of g is possible and leads to generalized hyper-g priors. The key advantages of using test-based BFs are computational efficiency and applicability to the Cox model.

Unfortunately it seems difficult to extend the approach to models with random effects. One interesting example of such models are the spline models used in Section 5.2 where the random effects correspond to the spline coefficients: We have to use data-based Bayes factors for these models; the computations are described in Sabanés Bové et al. (2012). In this approach, the random effects variances correspond to the spline smoothing parameters, which are part of the model space and thus not part of the parameter vector estimated in a model. Related work was presented by Cantoni and Hastie (2002), but only for the linear model. For unknown random effects variances that are part of the parameter vector, already the null distribution of the restricted likelihood ratio statistic is complicated, see Crainiceanu, Ruppert, Claeskens, and Wand (2005). In order to apply the test-based Bayes factors, we need in addition the asymp-

totic distribution under a sequence of local alternatives. Zhang and Lin (2008) tackle generalized linear mixed models, but their work does not extend to testing multiple random effects variances at once, which would be required to test each model against the null model. More research on the asymptotic distribution of the deviance under the null and local alternatives would thus be needed for the successful application of the proposed test-based BFs to mixed models.

Acknowledgments

The authors would like to thank Kerry L. Lee and Ewout W. Steyerberg for permitting to use the GUSTO-I data set.

Appendix

A. Proofs for Section 2.2

In Section 2.2.1 we state that the distribution of the deviance converges for $n \to \infty$ to a non-central chi-squared distribution with d_j degrees of freedom, where $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} \boldsymbol{\beta}_j$ is the non-centrality parameter. This is essentially proven by Davidson and Lever (1970), and we briefly show how their Theorem 1 applies here. In their notation the model is parametrized by $\theta = (\theta_1, \theta_2)$ with $\theta_1 = \boldsymbol{\beta}_j$ being the parameter of interest and $\theta_2 = \alpha$ being the nuisance parameter. We test the null hypothesis $\theta = (\theta_1^0, \theta_2)$ with $\theta_1^0 = 0$ and θ_2 unspecified. We assume the sequence of local alternatives $\theta^n = (\theta_1^n, \theta_2^*)$ with $\theta_{1k}^n = \delta_k / \sqrt{n}$, $k = 1, \ldots, d_j$, and $\theta_2^* = 0$. It follows that $\theta^n \to \theta^* = (0, \theta_2^*)$ for $n \to \infty$. Then Theorem 1 of Davidson and Lever (1970) states that for $n \to \infty$ the deviance converges in distribution to a non-central chi-squared distribution with d_j degrees of freedom and non-centrality parameter $\delta^\top \overline{C}_{11}(\theta^*)\delta$. Here $\overline{C}_{11}(\theta^*)$ is the inverse of the submatrix corresponding to θ_1 of the inverse expected Fisher information from one observation, evaluated at $\theta = \theta^*$. But we know that the expected Fisher information is block-diagonal for $\theta = \theta^*$, so $\overline{C}_{11}(\theta^*)$ is just the submatrix of the expected Fisher information from one observation. Moreover, for n observations we have $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j} =$

$n \cdot \overline{C}_{11}(\theta^*)$, and combined with $\delta = \sqrt{n}\boldsymbol{\beta}_j$ we obtain the non-centrality parameter $\lambda_j = \boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\boldsymbol{\beta}_j$.

In order to derive the prior distribution for λ_j based on the generalized g-prior (7) for $\boldsymbol{\beta}_j$ as stated in Section 2.2.2, first note that the generalized g-prior corresponds to

$$\tilde{\boldsymbol{\beta}}_j = \mathcal{I}^{1/2}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}/\sqrt{g}\boldsymbol{\beta}_j \sim \mathrm{N}_{d_j}(\mathbf{0}_{d_j}, \mathbf{I}_{d_j}),$$

where $\mathcal{I}^{1/2}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$ is the upper-triangular Cholesky root of $\mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}$. Hence we know that $\tilde{\boldsymbol{\beta}}_j^\top \tilde{\boldsymbol{\beta}}_j \sim \chi^2(d_j)$, which is identical to a $\mathrm{G}(d_j/2, 1/2)$ distribution. Expanding the quadratic form we find

$$\tilde{\boldsymbol{\beta}}_j^\top \tilde{\boldsymbol{\beta}}_j = 1/\sqrt{g}\boldsymbol{\beta}_j^\top \mathcal{I}^{\top/2}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\mathcal{I}^{1/2}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\boldsymbol{\beta}_j 1/\sqrt{g} = 1/g\boldsymbol{\beta}_j^\top \mathcal{I}_{\boldsymbol{\beta}_j,\boldsymbol{\beta}_j}\boldsymbol{\beta}_j = \lambda_j/g.$$

Since the second parameter of the gamma distribution is a rate parameter, we finally find $\lambda_j = g \cdot \lambda_j/g \sim \mathrm{G}(d_j/2, 1/(2g))$.

B. Proofs for Section 3.2

For the bounds mentioned in Section 3.2 usually the minimum Bayes factor in favour of the null hypothesis is considered, which is $\mathrm{mTBF}^{-1}_{j,0}$ in our notation. In order to simplify notation, we omit the model index j here. Let the P-value be $p = 1 - \mathrm{F}_{\chi^2(d)}(z)$, where $\mathrm{F}_{\chi^2(d)}$ is the cumulative distribution function of the chi-squared distribution with d degrees of freedom. Let $q = \Phi^{-1}(1 - p/2)$ be the corresponding quantile of the standard normal distribution with cumulative distribution function Φ. The proofs are adapted from Malaguerra (2012).

1. Let $d = 1$ and $z > d = 1$. We have $q^2 = z$ due to the fact that the squared standard normal distribution equals the chi-squared distribution with one degree of freedom. Hence we have

$$\begin{aligned}\mathrm{mTBF}^{-1}_{j,0} &= z^{1/2}\exp(-z/2)\exp(1/2)\\ &= q\exp(-q^2/2)\sqrt{e},\end{aligned}$$

   which is the required result from Berger and Sellke (1987).

2. Let $d = 2$ and $z > d = 2$. Due to the fact that $\mathrm{F}_{\chi^2(2)}(z) = 1 - \exp(-z/2)$ (Johnson, Kotz, and Balakrishnan, 1994, chapter 18), we have $p = \exp(-z/2)$ or $z = -2\log(p)$, such that $z > 2$ is equivalent to $p < 1/e$. Moreover,

$$\begin{aligned}\mathrm{mTBF}_{j,0}^{-1} &= (2/z)^{-1} \exp\left(-\frac{z-2}{2}\right) \\ &= \frac{z}{2}\, e\, \exp(-z/2) \\ &= -e\, p\, \log(p),\end{aligned}$$

which is the required result from Sellke et al. (2001).

3. The universal bound from Edwards et al. (1963) that we want to reach is $\exp(-q^2/2)$, so we have to show that for $d \to \infty$ and fixed P-value, the ratio of $\mathrm{mTBF}_{j,0}^{-1}$ and this universal bound is 1. With $d \to \infty$ we have that $(z-d)/\sqrt{2d} \overset{a}{\sim} \mathrm{N}(0,1)$, therefore due to the fixed P-value we have that

$$\frac{z-d}{\sqrt{2d}} \approx q.$$

Plugging this in (11) we obtain

$$\begin{aligned}\frac{\mathrm{mTBF}_{j,0}^{-1}}{\exp(-q^2/2)} &\approx \left(\frac{d}{\sqrt{2d}q + d}\right)^{-d/2} \exp\left(-\sqrt{\frac{d}{2}}q + q^2/2\right) \\ &= \exp\{-aq + a^2 \log(1 + q/a) + q^2/2\}\end{aligned}$$

with $a = \sqrt{d/2}$. Now for large d or a the term q/a is small, and hence we can do a second-order Taylor expansion of $\log(1+x)$ around $x = 0$, giving $\log(1+x) \approx x - x^2/2$. With this we have

$$\begin{aligned}\frac{\mathrm{mTBF}_{j,0}^{-1}}{\exp(-q^2/2)} &\approx \exp\left\{-aq + a^2\left(\frac{q}{a} - \frac{q^2}{2a^2}\right) + \frac{q^2}{2}\right\} \\ &= \exp(0) = 1,\end{aligned}$$

which proves the statement.

References

M. Banerjee. Likelihood ratio tests under local alternatives in regular semiparametric models. *Statistica Sinica*, 15(3):635–644, 2005.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004.

M. J. Bayarri, J. O. Berger, A. Forte, and G. García-Donato. Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics*, 40(3):1550–1577, 2012.

J. O. Berger and L. R. Pericchi. Objective Bayesian methods for model selection: introduction and comparison. In P. Lahiri, editor, *Model Selection*, volume 38 of *IMS Lecture Notes*, pages 135–207. Institute of Mathematical Statistics, Beachwood, OH, 2001.

J. O. Berger and T. Sellke. Testing a point null hypothesis - the irreconcilability of p-values and evidence. *Journal of the American Statistical Association*, 82(397):112–122, 1987.

E. Cantoni and T. Hastie. Degrees-of-freedom tests for smoothing splines. *Biometrika*, 89(2):251–263, 2002.

J. B. Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):311–354, 1983.

J. B. Copas. Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research*, 6(2):167–183, 1997.

C. Crainiceanu, D. Ruppert, G. Claeskens, and M. P. Wand. Exact likelihood ratio tests for penalised splines. *Biometrika*, 92(1):91–103, 2005.

W. Cui and E. I. George. Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008.

R. R. Davidson and W. E. Lever. The limiting distribution of the likelihood ratio statistic under a class of local alternatives. *Sankhyā: The Indian Journal of Statistics, Series A*, 32(2):209–224, 1970.

W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242, 1963.

M. Ennis, G. Hinton, D. Naylor, M. Revow, and R. Tibshirani. A comparison of statistical learning methods on the GUSTO database. *Statistics in Medicine*, 17(21):2501–2508, 1998.

C. Fernández, E. Ley, and M. F. J. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22(4):1947–1975, 1994.

S. Geisser. On prior distributions for binary trials. *The American Statistician*, 38(4): 244–247, 1984.

E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

S. N. Goodman. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130(12):1005–1013, 1999.

L. Held. A nomogram for P values. *BMC Medical Research Methodology*, 10(1):21, 2010.

N. L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

J. Hu and V. E. Johnson. Bayesian model selection using test statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(1):143–158, 2009.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, New York, 2 edition, 1994.

V. E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67(5):689–701, 2005.

V. E. Johnson. Properties of Bayes factors based on test statistics. *Scandinavian Journal of Statistics*, 35(2):354–368, 2008.

R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90 (431):928–934, 1995.

J. F. Lawless and K. Singhal. Efficient screening of nonnormal regression models. *Biometrics*, 34(2):318–327, 1978.

K. L. Lee, L. H. Woodlief, E. J. Topol, W. D. Weaver, A. Betriu, J. Col, M. Simoons, P. Aylward, F. Van de Werf, R. M. Califf, and for the GUSTO-I Investigators. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: results from an international trial of 41,021 patients. *Circulation*, 91(6):1659–1668, 1995.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481): 410–423, 2008.

D. V. Lindley. A statistical paradox. *Biometrika*, 44(1–2):187–192, 1957.

A. Malaguerra. Bayesian variable selection based on test statistics. Master's thesis, University of Zurich, 2012.

G. Marra and S. N. Wood. Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, 55(7):2372–2387, 2011.

S. A. Murphy and A. W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3):429–467, 1994.

D. Sabanés Bové and L. Held. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410, 2011a.

D. Sabanés Bové and L. Held. Bayesian fractional polynomials. *Statistics and Computing*, 21(3):309–324, 2011b.

D. Sabanés Bové and L. Held. On the Poisson approximation for hazard regression. *Biometrics*, 2013. to appear.

D. Sabanés Bové, L. Held, and G. Kauermann. Mixtures of g-priors for generalised additive model selection with penalised splines. Technical report, University of Zurich, 2012. URL http://arxiv.org/abs/1108.3520.

W. Sauerbrei, P. Royston, and H. Binder. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine*, 26(30):5512–5528, 2007.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619, 2010.

T. Sellke, M. J. Bayarri, and J. O. Berger. Calibration of p values for testing precise null hypotheses. *American Statistician*, 55(1):62–71, 2001.

E. Steyerberg. *Clinical Prediction Models*. Springer, New York, 2009.

E. W. Steyerberg, F. E. Harrell, G. J. Borsboom, M. Eijkemans, Y. Vergouwe, and J. F. Habbema. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8):774–781, 2001.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

J. C. Van Houwelingen and S. Le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 9(11):1303–1325, 1990.

C. T. Volinsky and A. E. Raftery. Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262, 2000.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 73(1):3–36, 2011.

Y. Yuan and V. E. Johnson. Bayesian hypothesis tests using nonparametric statistics. *Statistica Sinica*, 18:1185–1200, 2008.

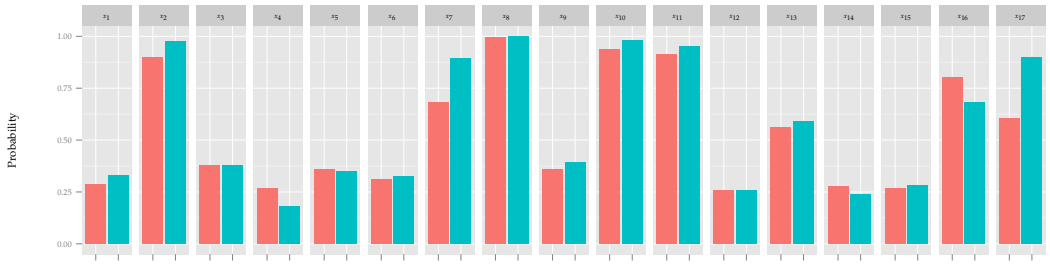
A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and*

Decision Techniques: Essays in Honor of Bruno de Finetti, volume 6 of *Studies in Bayesian Econometrics and Statistics*, chapter 5, pages 233–243. North-Holland, Amsterdam, 1986.

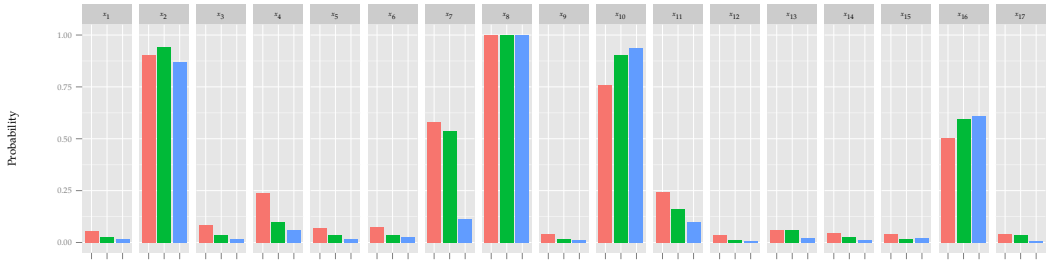
A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, pages 585–603, Valencia, 1980. University of Valencia Press.

D. Zhang and X. Lin. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In D. B. Dunson, editor, *Random Effect and Latent Variable Model Selection*, volume 192 of *Lecture Notes in Statistics*, pages 19–36. Springer, New York, 2008.

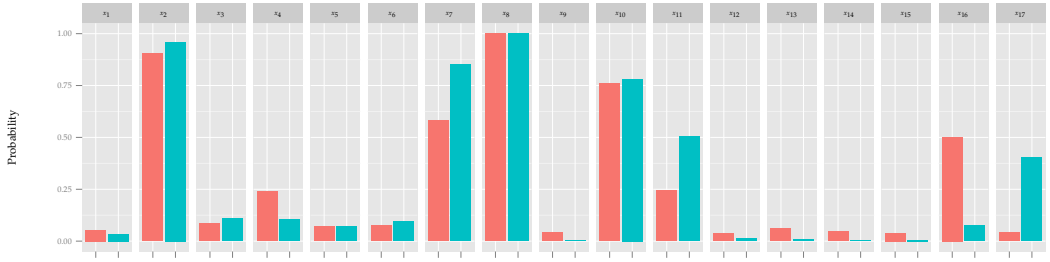
H. H. Zhang and W. Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.



(a) Plain variable selection: comparison of Cox with TBF (first bar, C) and Poisson approximation (second bar, P) inclusion probabilities.

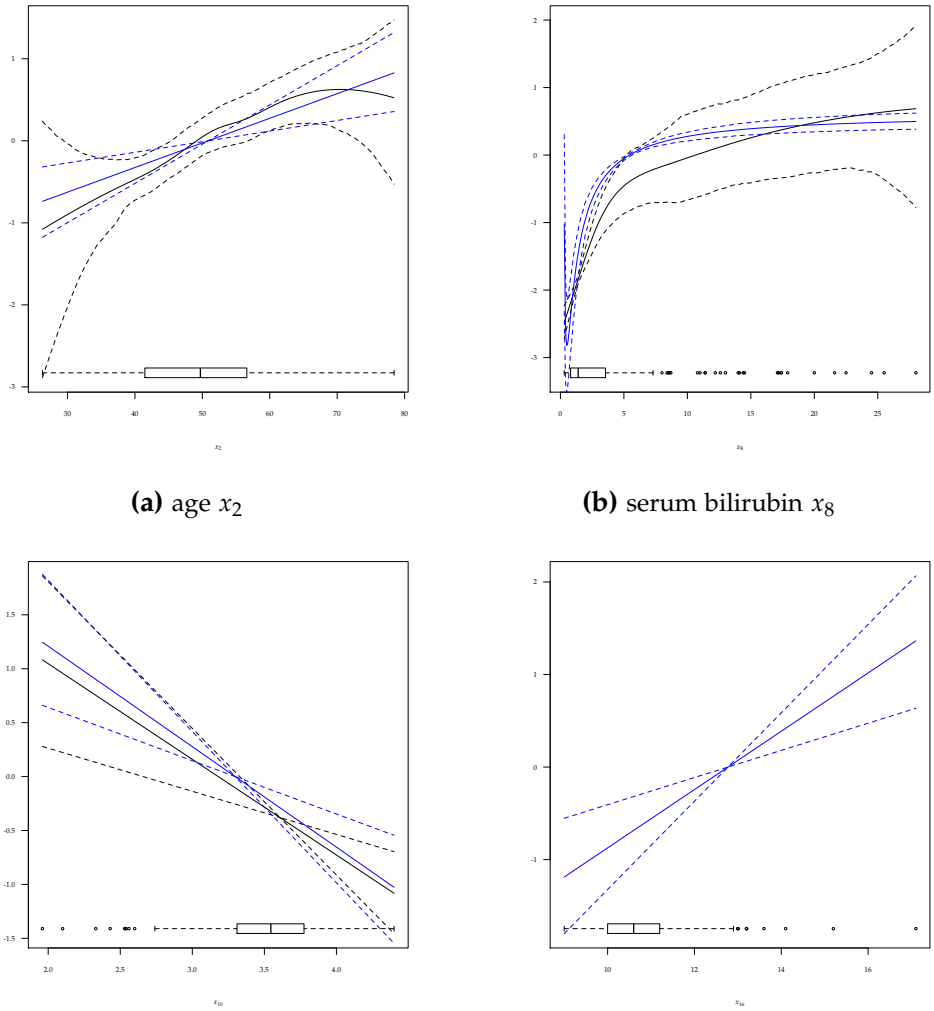


(b) Combined variable and FP function selection: comparison of Cox with TBF (first bar, CF), Poisson approximation with TBF (second bar, PT), and Poisson approximation with data-based BF (third bar, PF).



(c) Comparison of FP and spline modelling of nonlinear effects: Cox with TBF (first bar, CF) *versus* Poisson approximation with splines (second bar, PS).

Figure 5 – *PBC data: Inclusion probabilities, comparing the results from several approaches. In all cases the hyper-g / n_{obs} prior on g was used.*



(a) age x_2 **(b)** serum bilirubin x_8

(c) serum albumin x_{10} **(d)** standardised blood clotting time x_{16}

Figure 6 – *Comparison of MAP models from the FP (blue) and the splines approach (black). Means (solid lines) and pointwise (dashed lines) 95% credible intervals for the partial covariate effects (mean-centered) are given. Small boxplots at the bottom of the plots indicate data locations. The covariates x_7 (see the text) and x_{16} are missing in the FP and splines MAP models, respectively.*

APPENDIX I

Hyper-g priors for generalised additive model selection

Daniel Sabanés Bové, Leonhard Held & Göran Kauermann

Extended abstract published in the Proceedings of the 26^{th} *International Workshop on Statistical Modelling*, Valencia, Spain, 2011.

Hyper-g Priors for Generalised Additive Model Selection

Daniel Sabanés Bové[1], Leonhard Held[1] and Göran Kauermann[2]

[1] Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland, Email: daniel.sabanesbove@ifspm.uzh.ch, leonhard.held@ifspm.uzh.ch

[2] Centre for Statistics, Department of Economics and Business Administration, University Bielefeld, Postfach 300131, D-33501 Bielefeld, Germany, Email: gkauermann@uni-bielefeld.de

Abstract: We propose an automatic Bayesian approach to the selection of covariates and penalised splines transformations thereof in generalised additive models. Specification of a hyper-g prior for the model parameters and a multiplicity-correction prior for the models themselves is crucial for this task. We introduce the methodology in the normal model and illustrate it with an application to diabetes data. Extension to non-normal exponential families is finally discussed.

Keywords: Penalised splines; Bayesian variable selection; Shrinkage.

1 Introduction

Suppose we have p metrical covariates $x_1, \ldots, x_p$ and use the additive model $y = \beta_0 + \sum_{j=1}^{p} m_j(x_j) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. When x_j is included non-linearly in the model, we assume

$$m_j(x_j) = x_j\beta_j + \boldsymbol{Z}_j(x_j)^T\boldsymbol{u}_j$$

where $\boldsymbol{Z}_j(x_j)$ is the $K \times 1$ spline basis vector at position x_j and $\boldsymbol{u}_j \sim N(\boldsymbol{0}, \sigma^2\rho_j\boldsymbol{I})$ is the corresponding coefficients vector. In order to combine n observations, we stack these to the $n \times 1$ vector $\boldsymbol{x}_j$ and the $n \times K$ basis matrix $\boldsymbol{Z}_j$, both modified to be zero-centred and orthogonal to each other. We then translate the variance parameter ρ_j into the corresponding degree of freedom (Aerts, Claeskens and Wand, 2002, section 2.2)

$$d_j(\rho_j) = \mathrm{tr}\{(\boldsymbol{Z}_j^T\boldsymbol{Z}_j + \rho_j^{-1}\boldsymbol{I})^{-1}\boldsymbol{Z}_j^T\boldsymbol{Z}_j\} + 1 \in (1, K+1). \qquad (1)$$

A larger ρ_j (or a larger d_j) leads to a weaker penalty on the non-linear component of the function m_j. If x_j is excluded from or linearly included in the model we have $m_j(x_j) \equiv 0$ or $m_j(x_j) = x_j\beta_j$ and set $d_j = 0$ or

$d_j = 1$, respectively. Thus, the function m_j is exactly defined by d_j, which we may restrict to a finite set of values, say $d_j \in \{0, 1\} \cup \{2, 3, \ldots, K\}$. As default prior for the parameters β_0, $\boldsymbol{\beta} = (\beta_j : d_j \geq 1)$ and σ^2 in a given model specified via $\boldsymbol{d} = (d_1, \ldots, d_p)$,

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2 \sim \mathrm{N}(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\, \sigma^2\boldsymbol{I}) \tag{2}$$

with $\boldsymbol{X} = (\boldsymbol{x}_j : d_j \geq 1)$, $\boldsymbol{Z} = (\boldsymbol{Z}_j : d_j > 1)$ and $\boldsymbol{u} = (\boldsymbol{u}_j^T : d_j > 1)^T$, we propose the hyper-$g$ prior (Liang *et al.*, 2008) described in Section 2. For the models we propose a multiplicity-correction prior in Section 3. The methodology is applied to diabetes data in Section 4 and extended to generalised additive models in Section 5.

2 Hyper-g Priors for Additive Models

Integrating out the spline coefficients vector $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{D})$, where $\boldsymbol{D} = \mathrm{diag}\{\rho_j\boldsymbol{I} : d_j > 1\}$, from the conditional model (2) yields the marginal model

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}\left(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta},\, \sigma^2\boldsymbol{V}\right) \tag{3}$$

with $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T$ having Cholesky decomposition $\boldsymbol{V} = \boldsymbol{R}^T\boldsymbol{R}$. The transformed response vector $\tilde{\boldsymbol{y}} = \boldsymbol{R}^{-T}\boldsymbol{y}$ follows a linear model with similarly transformed design matrix $\tilde{\boldsymbol{X}}$ and diagonal covariance matrix $\sigma^2\boldsymbol{I}$. It turns out that we can use the hyper-g prior (Liang *et al.*, 2008) for this transformed model, *i. e.* a locally uniform prior $p(\beta_0) \propto 1$ on the intercept, Jeffreys' prior $p(\sigma^2) \propto (\sigma^2)^{-1}$ on the variance and the g-prior (Zellner, 1986)

$$\boldsymbol{\beta} \,|\, g, \sigma^2 \sim \mathrm{N}\left(\mathbf{0},\, g\sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\right) \tag{4}$$

on the coefficients are combined with a uniform prior on the shrinkage coefficient $g/(1+g)$. Note that $\sigma^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$ is the Fisher information matrix of $\boldsymbol{\beta}$ in the marginal model (3). The hyper-g prior leads to a closed form of the marginal likelihood, which we can compute on the original response scale via the change of variables formula:

$$p(\boldsymbol{y} \,|\, \boldsymbol{d}) \propto \|\tilde{\boldsymbol{y}} - \bar{\tilde{\boldsymbol{y}}}\|^{-(n-1)}\,(l_{\boldsymbol{d}} + 2)^{-1}\,{}_2F_1\left(\frac{n-1}{2}; 1; \frac{l_{\boldsymbol{d}}+4}{2}; \tilde{R}^2\right)|\boldsymbol{R}|^{-1},$$

where $l_{\boldsymbol{d}}$ is the dimension of $\boldsymbol{\beta}$, ${}_2F_1$ is the Gaussian hypergeometric function and $\tilde{R}^2$ is the classical coefficient of determination in model (3).

3 Model Prior

We propose a prior $p(\boldsymbol{d})$ on the model space which explicitly corrects for the multiplicity of testing inherent in the simultaneous analysis of many

TABLE 1. Marginal posterior probabilities (x_1: age, x_2: systolic blood pressure, x_3: cholesterol/HLD ratio, x_4: BMI, x_5: waist/hip ratio, x_6: gender).

	x_1	x_2	x_3	x_4	x_5	x_6
not included ($d_j = 0$)	0.00	0.65	0.00	0.14	0.50	0.65
linear ($d_j = 1$)	0.71	0.33	0.93	0.81	0.48	0.35
non-linear ($d_j > 1$)	0.29	0.03	0.07	0.05	0.02	—

covariates (see Scott and Berger, 2010): *A priori*, the number of covariates included in the model ($l_{\boldsymbol{d}}$) is uniformly distributed on $\{0, 1, \ldots, p\}$. Then the number of non-linearly included covariates ($s_{\boldsymbol{d}}$) is uniformly distributed on $\{0, 1, \ldots, l_{\boldsymbol{d}}\}$. The respective choice of the $l_{\boldsymbol{d}}$ and $s_{\boldsymbol{d}}$ covariates is uniformly distributed on all possible configurations. Finally, the degrees of freedom of the non-linearly modelled covariates are independent and uniformly distributed on $\{2, 3, \ldots, K\}$. Altogether, this gives

$$1/p(\boldsymbol{d}) = \binom{p}{l_{\boldsymbol{d}}}(p+1)\binom{l_{\boldsymbol{d}}}{s_{\boldsymbol{d}}}(l_{\boldsymbol{d}}+1)(K-1)^{s_{\boldsymbol{d}}}$$

and leads to marginal prior probabilities $\Pr(d_j = 0) = 1/2$, $\Pr(d_j = 1) = \Pr(d_j > 1) = 1/4$.

4 Application

We illustrate our modelling approach with the diabetes data from Harrell (2001). We study the association of (the negative reciprocal of) glycosolated haemoglobin of $n = 377$ study participants with the continuous covariates age (in years), systolic blood pressure (in mmHg), cholesterol/HDL ratio, body mass index (BMI, in kg/m^2) and waist/hip ratio as well as the binary covariate gender. As the computational complexity is quadratic in the spline basis dimension K, we want to use splines with few quantile-based knots. Therefore, we choose cubic O'Sullivan splines (Wand and Ormerod, 2008). Here, we get basis matrices $\boldsymbol{Z}_j$ with $K = 9$ columns from 7 knots. The exhaustive evaluation of the posterior model probabilities $p(\boldsymbol{d} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{d})p(\boldsymbol{d})$ of all $(K+1)^5 \cdot 2 = 200\,000$ models takes only 585 seconds due to an efficient `C++` implementation which is available in an `R`-package from the first author. In Table 1 the marginal posterior probabilities for linear and non-linear inclusion of the six covariates are shown. There is strong evidence for linear inclusion of cholesterol/HDL ratio and BMI, while the posterior probability for inclusion of systolic blood pressure or gender is only 35%. There is overwhelming evidence for (non-linear) inclusion of age, and the posterior odds for (linear) inclusion of waist/hip ratio are around 1. The *maximum a posteriori* model includes age, cholesterol/HDL ratio and BMI all linearly. Note that these are the covariates

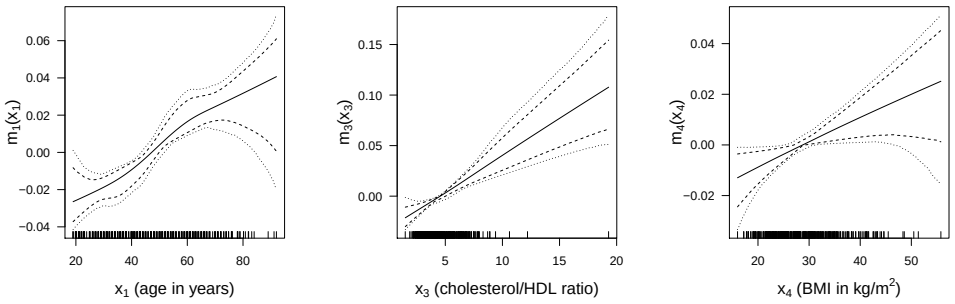


FIGURE 1. Estimated covariate effects in the median probability model average, based on 10 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals as well as positions of data points (ticks above x-axes) are shown.

which have inclusion probabilities larger than 50%, thus defining the set of median probability models (Barbieri and Berger, 2004) $\boldsymbol{d}$ with $d_1, d_3, d_4 \geq 1$ and $d_2 = d_5 = d_6 = 0$. Figure 1 shows the estimated covariate effects from the resulting model average. While the age effect is slightly non-linear (with 38% probability in the median probability models), both other covariates have essentially linear effect estimates.

5 Extension to Generalised Additive Models

Now we assume more generally that the covariate effects $m_j(x_j)$ enter additively into the linear predictor $\eta = \beta_0 + \sum_{j=1}^{p} m_j(x_j)$ of an exponential family distribution with canonical parameter θ, mean $\mathrm{E}(y) = h(\eta) = db(\theta)/d\theta$ and variance $\mathrm{Var}(y) = \phi/w \cdot v(\mu) = \phi/w \cdot d^2b(\theta)/d\theta^2$ (see McCullagh and Nelder, 1989). We restrict our attention to non-normal distributions with fixed dispersion ϕ (as $\phi = 1$ for the Bernoulli and Poisson distribution) and known weight w. For n observations, the linear predictor vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ is $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$, where $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$, and the likelihood is

$$p(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left\{ \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} \right\}. \tag{5}$$

A reasonable generalisation of (1) is (see Ruppert, Wand and Carroll, 2009, section 11.4)

$$d_j(\rho_j) = \mathrm{tr}\{(\boldsymbol{Z}_j^T \widehat{\boldsymbol{W}} \boldsymbol{Z}_j + \rho_j^{-1}\boldsymbol{I})^{-1} \boldsymbol{Z}_j^T \widehat{\boldsymbol{W}} \boldsymbol{Z}_j\} + 1, \tag{6}$$

which uses a fixed weight matrix $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\mathbf{1}\widehat{\beta}_0)$ for all models, where $\boldsymbol{W}(\boldsymbol{\eta}) = \mathrm{diag}\{(dh(\eta_i)/d\eta)^2 v(h(\eta_i))^{-1}\phi^{-1}w_i\}_{i=1}^n$ is the usual generalised linear model (GLM) weight matrix and $\widehat{\beta}_0$ is the intercept estimate from the null model. Therefore, we now arrange $\mathbf{1}$, $\boldsymbol{x}_j$ and the columns of $\boldsymbol{Z}_j$ to

be orthogonal with respect to the inner product in terms of $\widehat{\boldsymbol{W}}$, so that (6) correctly captures the degrees of freedom associated with the non-linear part of m_j.
In order to derive a generalised g-prior for $\boldsymbol{\beta}$, we will use the iterative weighted least squares (IWLS) approximation to (5) to come back to a normal model and then derive the resulting g-prior (4). So let

$$\boldsymbol{z}_0 = \boldsymbol{\eta}_0 + \mathrm{diag}\{dh(\boldsymbol{\eta}_0)/d\boldsymbol{\eta}\}^{-1}(\boldsymbol{y} - h(\boldsymbol{\eta}_0))$$

be the adjusted response vector resulting from a first-order approximation to $h^{-1}(\boldsymbol{y})$ around $h(\boldsymbol{\eta}_0)$, such that

$$\boldsymbol{z}_0 \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u} \sim \mathrm{N}\big(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\, \boldsymbol{W}(\boldsymbol{\eta}_0)^{-1}\big)$$

is the working normal model. This can be rewritten to

$$\tilde{\boldsymbol{z}}_0 \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u} \sim \mathrm{N}(\tilde{\mathbf{1}}\beta_0 + \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{u},\, \boldsymbol{I}) \tag{7}$$

by setting $\tilde{\boldsymbol{z}}_0 = \boldsymbol{W}(\boldsymbol{\eta}_0)^{1/2}\boldsymbol{z}_0$ *etc.* Since (7) is analogous to (2), our proposal for a generalised g-prior is

$$\boldsymbol{\beta} \,|\, g \sim \mathrm{N}(\mathbf{0}, g\boldsymbol{J}^{-1}), \tag{8}$$

where $\boldsymbol{J}$ is the Fisher information for $\boldsymbol{\beta}$ in (7) with $\boldsymbol{\eta}_0 = \mathbf{0}$:

$$\begin{aligned}\boldsymbol{J} &= \tilde{\boldsymbol{X}}^T(\boldsymbol{I} + \tilde{\boldsymbol{Z}}\boldsymbol{D}\tilde{\boldsymbol{Z}}^T)^{-1}\tilde{\boldsymbol{X}} \\ &= \boldsymbol{X}^T\boldsymbol{W}_0^{1/2}(\boldsymbol{I} + \boldsymbol{W}_0^{1/2}\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{W}_0^{1/2})^{-1}\boldsymbol{W}_0^{1/2}\boldsymbol{X},\end{aligned}$$

abbreviating $\boldsymbol{W}_0 = \boldsymbol{W}(\mathbf{0})$. Note that this prior directly generalises the prior proposed by Sabanés Bové and Held (2011) for GLMs, to which it reduces when there are no spline effects in the model.
The generalised hyper-g prior then consists of the improper prior $p(\beta_0) \propto 1$ on the intercept β_0, the g-prior (8) on the linear effects vector $\boldsymbol{\beta}$, the penalty prior $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$ on the spline coefficients vector $\boldsymbol{u}$ and some proper hyper-prior $p(g)$ on the hyper-parameter g in the g-prior. For the implementation of posterior inference we can easily extend the approach of Sabanés Bové and Held (2011, section 3). Let $\boldsymbol{X}_a = (\mathbf{1}, \boldsymbol{X}, \boldsymbol{Z})$ and $\boldsymbol{\beta}_a = (\beta_0, \boldsymbol{\beta}^T, \boldsymbol{u}^T)^T$, such that $\boldsymbol{\eta} = \boldsymbol{X}_a\boldsymbol{\beta}_a$. The prior for $\boldsymbol{\beta}_a$ conditional on g has Gaussian form with mean zero and singular precision $\boldsymbol{R}_a = \mathrm{diag}\{0, g^{-1}\boldsymbol{J}(\mathbf{0}), \boldsymbol{D}^{-1}\}$. Thus, the Laplace approximation of $p(\boldsymbol{y} \,|\, g, \boldsymbol{d})$, which is based on a Gaussian approximation to the conditional posterior $p(\boldsymbol{\beta}_a \,|\, \boldsymbol{y}, g)$, can be obtained by the Bayesian IWLS algorithm (West, 1985). Afterwards, the marginal likelihood

$$p(\boldsymbol{y} \,|\, \boldsymbol{d}) = \int_0^\infty p(\boldsymbol{y} \,|\, g, \boldsymbol{d})p(g)\, dg,$$

can be approximated by numerical integration of the Laplace approximation $\tilde{p}(\boldsymbol{y} \mid g, \boldsymbol{d})$. Note that this strategy of integrated Laplace approximations was proposed more generally by Rue, Martino and Chopin (2009). Finally, for sampling from the posterior of $\boldsymbol{\beta}_a$ and g in a specific model $\boldsymbol{d}$ we can use a tuning-free Metropolis-Hastings algorithm.

References

Aerts, M., Claeskens, G., and Wand, M.P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, **103**, 455-470.

Barbieri, M.M., and Berger, J.O. (2004). Optimal predictive model selection. *Annals of Statistics*, **32**, 870-897.

Harrell, Jr., F.E. (2001). *Regression Modeling Strategies*. New York: Springer.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410-423.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **71**, 319-392.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Sabanés Bové, D., and Held, L. (2011). Hyper-g Priors for Generalized Linear Models. *Bayesian Analysis*, **6**, forthcoming article. URL: `http://ba.stat.cmu.edu/abstracts/Sabanes.php`

Scott, J.G., and Berger, J.O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, **38**, 2587-2619.

Wand, M.P., and Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, **50**, 179-198.

West, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions. In: *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, 531-558. Amsterdam: North-Holland.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233-243. Amsterdam: North-Holland.

APPENDIX II

Software manual

Introduction to the R-packages

Daniel Sabanés Bové*

September 2013

Institute of Social and Preventive Medicine

Division of Biostatistics

This appendix shall serve as an introduction to the R-packages that I wrote for the implementation of the methods of my thesis. The two main R-packages are `glmBfp` (encompasses Papers I and IV) and `hypergsplines` (for Papers II and III). Both packages are hosted on R-Forge and can be installed with the following command:

```
install.packages(c("glmBfp", "hypergsplines"), repos = "http://R-Forge.R-project.org")
```

Note that for fractional polynomial model selection with Gaussian response (Sabanés Bové and Held, 2011b), the R-package `bfp` is available on `CRAN`.

In addition, I wrote the R-package `appell`, which makes Fortran code by Colavecchia and Gasaneo (2004) accessible for computing the Appell's F1 hypergeometric function. Moreover, the hypergeometric function with complex arguments is computed with Fortran code by Michel and Stoitsov (2008) or with Fortran code by Forrey (1997). The package is required for computing the log marginal likelihood in case when the hyper-g/n prior is used in additive models with the `hypergsplines` package. The package `appell` is available on `CRAN`.

*E-mail: daniel.sabanesbove@ifspm.uzh.ch

1 Logistic Regression

First, we would like to illustrate the use of the R-packages for logistic regression. We will have a look at the Pima Indians diabetes data (Ripley, 1996; Frank and Asuncion, 2010), which is available in the package MASS (Venables and Ripley, 2002):

```
library(MASS)
pima <- rbind(Pima.tr, Pima.te)
pima$hasDiabetes <- as.numeric(pima$type == "Yes")
pima.nObs <- nrow(pima)
```

For $n = 532$ women of Pima Indian heritage, seven possible predictors for the presence of diabetes are recorded. We would like to investigate with logistic regression, which of them are relevant, and what form the statistical association has—is it a linear effect, or rather a nonlinear effect?

1.1 Fractional Polynomials

In this section we will model possible nonlinear effects with the fractional polynomials (FPs), using the R-package glmBfp. We are going to use the generalised hyper-g priors for GLMs (Sabanés Bové and Held, 2011a). They are automatic, and only require the specification of the hyperprior for the prior covariance factor g.

Hyperprior on g One possible choice is the Zellner-Siow hyperprior which assumes $g \sim \text{IG}(1/2, n/2)$:

```
library(glmBfp)
## define the prior distributions for g which we are going to
## use:
prior <- InvGammaGPrior(a = 1/2, b = pima.nObs/2)

## Warning:  density must be proper and normalized:  (numerical) integral is 1.00000013575927
```

This corresponds to the F1 prior in Sabanés Bové and Held (2011a). Note that a normalisation warning is printed, which can be ignored because we know that the density is properly normalised. Similarly, the hyper-g prior (Liang, Paulo, Molina, Clyde, and Berger, 2008) (with default parameter $a = 4$ corresponding to a uniform prior on the

shrinkage factor $t = g/(g+1)$) can be used with the `HypergPrior` function. For all other hyperpriors, there is no special constructor function. One example is the hyper-g/n prior. To use such a prior, you can instead directly specify the log prior density function with `CustomGPrior`, as follows:

```
## You may also use the hyper-g/n prior:
prior.f2 <- CustomGPrior(logDens = function(g) {
    -log(pima.nObs) - 2 * log(1 + g/pima.nObs)
})
```

Empirical Bayes estimation of g Alternatively, you can use local empirical Bayes estimation of g by specifying `empiricalBayes=TRUE` in the `glmBayesMfp` call below. Then it does not matter which prior on g you specify, because it will not be used for the computations. Only for technical reasons you still have to specify any one of the hyperpriors.

You can even use a fixed value of g, using the `fixedg` option of `glmBayesMfp`. For example, in order to use a unit information prior, you can here set `fixedg=pima.nObs`, which sets $g = n$.

Model specification and prior For specifying the possible models, we use a formula like

```
formula.pima <- type ~ bfp(npreg) + bfp(glu) + bfp(bp) + bfp(skin) +
    bfp(bmi) + bfp(ped) + bfp(age)
```

similarly to other regression functions in `R`. With the `bfp` function, we specify which covariates should be modelled with the FPs. Note that you can also restrict the maximum degree of the FPs and avoid preliminary scaling of the covariates with this function, see the help page `?bfp` for more details.

If you have covariates for which you would only like to do selection without FP transformation, you can instead of `bfp` use the function `uc` ("uncertain but fixed form"). This allows to do variable selection without combined function selection. Moreover, if you have a factor variable (say `var`) with multiple levels, you can still use `uc(var)` in the formula. If the covariate is selected to be included, then all dummy variables representing the different levels will be included in the model.

Moreover, we have to decide on the model prior, and in this example we use the `sparse` type which was also used in Sabanés Bové and Held (2011a). The alternatives are `flat`, which assumes that all models have the same prior probability, and `dependent`, which is the multiplicity-correcting model prior described in Paper IV, see also Section 2.3 in the Introduction. The model prior option is passed together in a list with the hyperprior for g in the argument `priorSpecs` of `glmBayesMfp`:

```
priors.pima <- list(gPrior = prior, modelPrior = "sparse")
```

Stochastic model search Next, we will do a stochastic search on the (very large) model space to find "good" models. We use a `chainlength` of 100, which is very small but enough for illustration purposes (in practice one should use at least 10 000 as a rule of thumb), and save all models because we set `nModels` to a value that is larger than `chainlength` (in general `nModels` is the number of models which are saved from all visited models). The stochastic search is invoked by `method="sampling"`, and the alternative `"exhaustive"` computes all models, *i. e.* it exhaustively explores the model space. See Section 3 in the Introduction for more details on model search algorithms.

Finally, we decide that we do not want to use OpenMP acceleration (this would parallelise loops over all observations on all cores of your processor) and that we want to do higher order correction for the Laplace approximations (*cp.* Paper I). The progress could be monitored interactively with the option `verbose=TRUE`, which is however not useful in a static document like this manual. In order to be able to reproduce the analysis, it is advisable to set a seed for the random number generator before starting the stochastic search.

```
set.seed(102)
time.pima <- system.time(models.pima <- glmBayesMfp(formula.pima,
    data = pima, family = binomial("logit"), priorSpecs = priors.pima,
    nModels = 1000L, chainlength = 10L, method = "sampling",
    useOpenMP = FALSE, higherOrderCorrection = TRUE, verbose = FALSE))
time.pima

##    user  system elapsed
##   1.048   0.008   1.658
```

```
attr(models.pima, "numVisited")

## [1] 9
```

We see that the search took 2 seconds, and 9 models were encountered. Now, if we want to have a table of the saved models, ordered according to their posterior probability, and including the log marginal likelihood, the log prior probability, the powers for every covariate, and the frequency that the sampler visited that model:

```
table.pima <- as.data.frame(models.pima, freq = TRUE)
table.pima

##   posterior frequency logMargLik logPrior age bmi bp glu npreg ped
## 1 5.751e-01       0.1     -265.8   -11.85              1
## 2 4.037e-01       0.4     -266.2   -11.85              2
## 3 7.967e-03       0.3     -270.1   -11.85              2
## 4 6.911e-03       0.0     -268.2   -13.93          0   2
## 5 5.925e-03       0.1     -272.5    -9.77              2
## 6 3.275e-04       0.1     -273.3   -11.85          3   2
## 7 2.148e-20       0.0     -310.5   -11.85   2
## 8 2.194e-23       0.0     -317.4   -11.85                   2
## 9 2.562e-31       0.0     -339.9    -7.69
##   skin
## 1 -0.5
## 2 -0.5
## 3    3
## 4 -0.5
## 5
## 6
## 7 -0.5
## 8    3
## 9
```

Note that while `frequency` refers to the frequency of the models in the sampling chain, thus providing a Monte Carlo estimate of the posterior model probabilities, `posterior` refers to the renormalised posterior model probabilities. The latter has the advantage that ratios of posterior probabilities between any two models are exact, while the former is unbiased (but obviously has larger variance). See Section 3,

page 9, in the Introduction for more discussion and García-Donato and Martinez-Beneito (2013) for a thorough comparison of both strategies.

Inclusion probabilities The estimated marginal inclusion probabilities for all covariates are also saved:

```
round(attr(models.pima, "inclusionProbs"), 3)

##   age   bmi    bp   glu npreg   ped  skin
## 0.000 0.000 0.007 1.000 0.000 0.000 0.994
```

These probabilities are based on the renormalised model probabilities of all models encountered during the stochastic model search. In this toy example, we saved all models, so there is no difference to

```
round(inclusionProbs(models.pima), 3)

##   age   bmi    bp   glu npreg   ped  skin
## 0.000 0.000 0.007 1.000 0.000 0.000 0.994
```

which gives the probabilities based on all saved models. We can also obtain the results based on the model frequencies:

```
inclusionProbs(models.pima, postProbs = posteriors(models.pima,
    type = "sampling"))

##   age   bmi    bp   glu npreg   ped  skin
##   0.0   0.0   0.1   1.0   0.0   0.0   0.8
```

Sampling model parameters If we now want to look at the estimated covariate effects in the estimated *maximum a posteriori* (MAP) model which has the configuration given in the first row and the last seven columns of `table.pima`, then we first need to generate parameter samples from that model:

```
## MCMC settings
mcmcOptions <- McmcOptions(iterations = 1000L, burnin = 100L,
    step = 2L)
```

```
## get samples from the MAP model
set.seed(634)
mapSamples <- sampleGlm(models.pima[1L], mcmc = mcmcOptions,
    useOpenMP = FALSE, verbose = FALSE)
mapSamples$acceptanceRatio

## [1] 0.865
```

With the function `McmcOptions`, we have defined an S4 object of Markov chain Monte Carlo (MCMC) settings, comprising the number of iterations, the length of the burn-in, the thinning step (here we save every second iteration). Here the acceptance rate was 0.86, which is quite high and thus indicates that the sampling worked well. Note that you can also get predictive samples for new data points via the `newdata` option of `sampleGlm`.

The result `mapSamples` has the following structure:

```
str(mapSamples)

## List of 5
## $ tbf : logi FALSE
## $ acceptanceRatio: num 0.865
## $ logMargLik :List of 5
## ..$ estimate : Named num -266
## .. ..- attr(*, "names")= chr "numeratorTerms"
## ..$ standardError : num [1, 1] 0.0138
## ..$ numeratorTerms : num [1:450] 0.519 0.701 0.652 0.603 0.637 ...
## ..$ denominatorTerms : num [1:450] 0.867 1 0.971 1 1 ...
## ..$ highDensityPointLogUnPosterior: num -266
## $ coefficients : num [1:3, 1:450] -0.978 4.458 -4.873 -0.645 3.929
##    ...
## $ samples :Formal class 'GlmBayesMfpSamples' [package "glmBfp"] with
##    8 slots
## .. ..@ fitted : num [1:532, 1:450] -2.44 2.65 -2.33 1.65 -1.67 ...
## .. .. ..- attr(*, "dimnames")=List of 2
## .. .. .. ..$ : chr [1:532] "1" "2" "3" "4" ...
## .. .. .. ..$ : NULL
## .. ..@ predictions : logi[0 , 0 ]
## .. ..@ fixed : num [1:450] -0.978 -0.645 -0.896 -0.967 -0.901 ...
## .. ..@ z : num [1:450] 4.66 5.52 6.46 7.49 6.32 ...
```

```
## .. ..@ bfpCurves :List of 2
## .. .. ..$ glu : num [1:327, 1:450] -3.23 -3.2 -3.19 -3.17 -3.14 ...
## .. .. .. ..- attr(*, "scaledGrid")= num [1:327, 1] 0.56 0.567 0.57
##    0.574 0.581 ...
## .. .. .. .. ..- attr(*, "dimnames")=List of 2
## .. .. .. .. .. ..$ : NULL
## .. .. .. .. .. ..$ : chr "glu"
## .. .. .. ..- attr(*, "whereObsVals")= int [1:532] 62 318 40 251 113
##    88 55 313 197 163 ...
## .. .. ..$ skin: num [1:251, 1:450] -3.28 -3.09 -2.93 -2.9 -2.78 ...
## .. .. .. ..- attr(*, "scaledGrid")= num [1:251, 1] 0.7 0.746 0.791
##    0.8 0.837 ...
## .. .. .. .. ..- attr(*, "dimnames")=List of 2
## .. .. .. .. .. ..$ : NULL
## .. .. .. .. .. ..$ : chr "skin"
## .. .. .. ..- attr(*, "whereObsVals")= int [1:532] 67 83 108 115 57 63
##    76 28 25 95 ...
## .. ..@ ucCoefs : list()
## .. ..@ shiftScaleMax: num [1:7, 1:4] 0 0 0 0 1 0 0 10 100 100 ...
## .. .. ..- attr(*, "dimnames")=List of 2
## .. .. .. ..$ : chr [1:7] "age" "bmi" "bp" "glu" ...
## .. .. .. ..$ : chr [1:4] "shift" "scale" "maxDegree" "cardPowerset"
## .. ..@ nSamples : int 450
```

It is a list with the `acceptanceRatio` of the Metropolis-Hastings proposals, an MCMC estimate for the log marginal likelihood including an associated standard error (`logMargLik`), the `coefficients` samples of the model, and an S4 object `samples`. This S4 object includes the `fitted` samples on the linear predictor scale (in our case on the log odds ratio scale), possibly `predictions` samples, samples of the intercept (`fixed`), samples of $z = \log(g)$ (`z`), samples of the FP curves (`bfpCurves`), coefficients of factor or untransformed variables (`ucCoefs`), the shifts and scales applied to the original covariates for numerical reasons (`shiftScaleMax`) and the number of samples (`nSamples`). You can read more details about the results on the help page by typing `?"GlmBayesMfpSamples-class"` in `R`.

If we wanted to get posterior fitted values on the probability scale, we can use the following code:

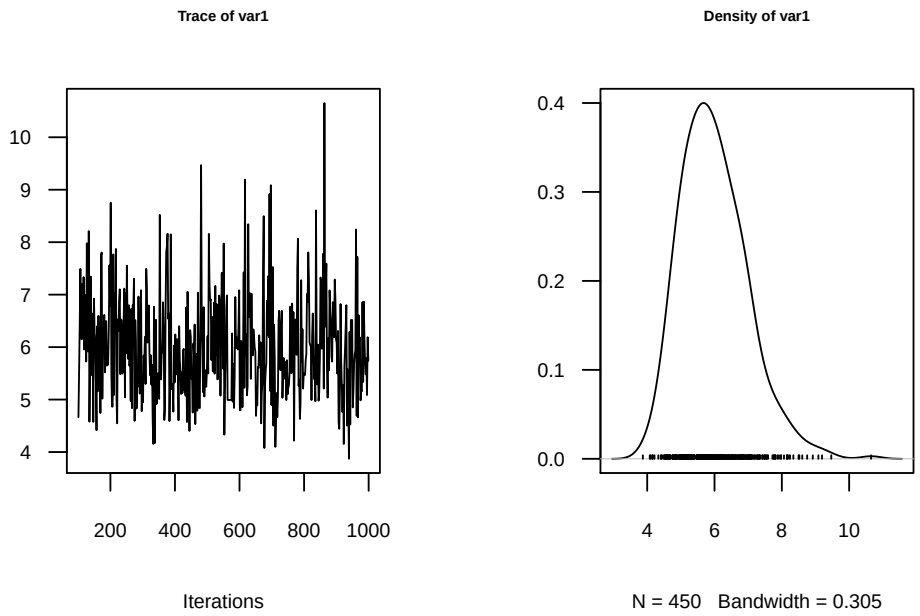
```
mapFit <- rowMeans(plogis(mapSamples$samples@fitted))
head(mapFit)

##      1      2      3      4      5      6
## 0.1017 0.8955 0.1093 0.7842 0.1792 0.1410
```

We can also analyse the MCMC output in greater detail using the coda package (Plummer, Best, Cowles, and Vines, 2006):

```
library(coda)
coefMcmc <- mcmc(data = t(mapSamples$coefficients), start = mcmcOptions@burnin +
    1, thin = mcmcOptions@step)
str(coefMcmc)

## mcmc [1:450, 1:3] -0.978 -0.645 -0.896 -0.967 -0.901 ...
## - attr(*, "mcpar")= num [1:3] 101 999 2

## standard summary table for the coefficients:
summary(coefMcmc)

##
## Iterations = 101:999
## Thinning interval = 2
## Number of chains = 1
## Sample size per chain = 450
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##        Mean    SD Naive SE Time-series SE
## [1,] -0.935 0.122  0.00574         0.0065
## [2,]  3.842 0.409  0.01928         0.0274
## [3,] -4.024 1.089  0.05135         0.0618
##
## 2. Quantiles for each variable:
##
##       2.5%   25%    50%    75%  97.5%
## var1 -1.19 -1.01 -0.932 -0.848 -0.715
## var2  3.03  3.54  3.849  4.119  4.701
## var3 -6.37 -4.74 -4.051 -3.242 -2.067

## etc., e.g.  autocorr(coefMcmc) plot(coefMcmc)
```

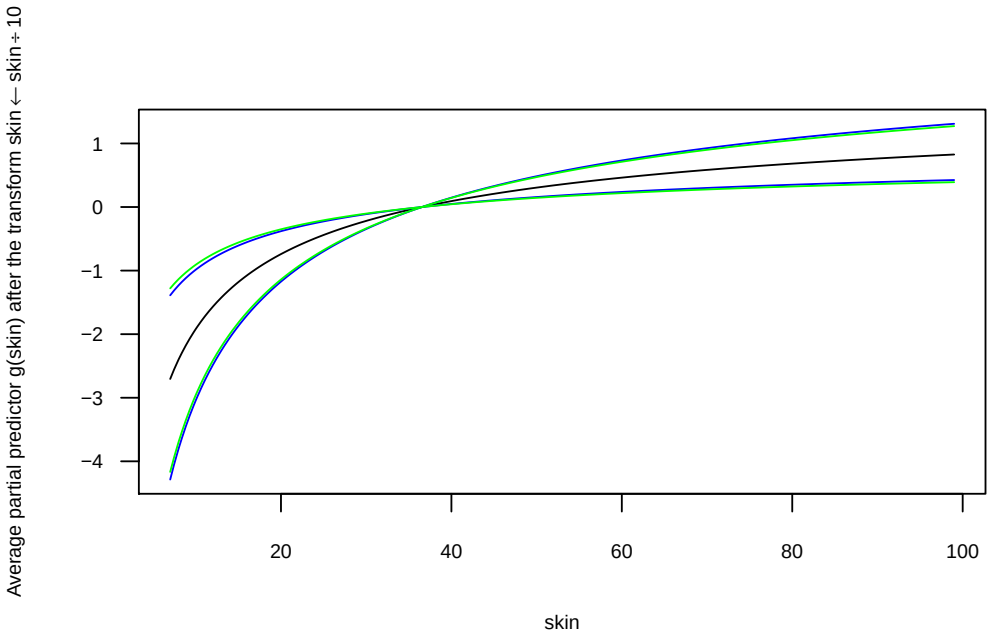
For the samples of $z = \log(g)$, we can obtain a trace and a density plot as follows:

```
## samples of z:
zMcmc <- mcmc(data = mapSamples$samples@z, start = mcmcOptions@burnin +
    1, thin = mcmcOptions@step)
plot(zMcmc)
```



Curve estimates Now we can use the samples to plot the estimated effects of the MAP model covariates, with the `plotCurveEstimate` function. For example:

```
plotCurveEstimate(termName = "skin", samples = mapSamples$samples)
```



The mean curve is plotted in black. Pointwise 95% credible intervals are plotted in blue, and a simultaneous 95% credible band is plotted in green. The credible levels can be customized with the options `plevel` and `slevel`, respectively.

Model averaging Bayesian model averaging (BMA) works in principle similarly to sampling from a single model, but multiple model configurations are supplied and their respective log posterior probabilities. For example, if we wanted to average the top three models found, we would do the following:

```
set.seed(312)
bmaSamples <- sampleBma(models.pima[1:3], mcmc = mcmcOptions,
    useOpenMP = FALSE, nMargLikSamples = 1000, verbose = FALSE)
## look at the list element names:
names(bmaSamples)

## [1] "modelData" "samples"

## now we can see how close the MCMC estimates
## ('margLikEstimate') are to the ILA estimates ('logMargLik')
## of the log marginal likelihood:
bmaSamples$modelData[, c("logMargLik", "margLikEstimate")]
```

```
##   logMargLik margLikEstimate
## 1     -265.8          -265.8
## 2     -266.2          -266.2
## 3     -270.1          -270.1

## the 'samples' list is again of class 'GlmBayesMfpSamples':
class(bmaSamples$samples)

## [1] "GlmBayesMfpSamples"
## attr(,"package")
## [1] "glmBfp"
```

Then internally, first the models are sampled, and for each sampled model so many samples are drawn as determined by the model frequency in the model average sample. The result is a list with two elements: `modelData` is similar to the `table.pima`, and contains in addition to that the BMA probability and frequency in the sample, the MCMC acceptance ratios (which should be high). On the second element `samples`, which is again of class `GlmBayesMfpSamples`, the above presented functions can again be applied (e.g. `plotCurveEstimate`).

Test-based Bayes factors You can use the test-based Bayes factor (TBF) methodology from Paper IV by specifying `tbf=TRUE` in the `glmBayesMfp` call.

In order to use the conjugate incomplete inverse-gamma prior for g, you can use the function `IncInvGammaGPrior`, which takes the a and b parameters. If you specify *e.g.* the hyper-g prior using `IncInvGammaGPrior(a=1, b=0)`, then the Bayes factors will be computed using the closed form expression given in the paper. If you use instead `HypergPrior(a=4)`, then numerical integration will be used to compute the Bayes factors.

Basically all functions described above work the same way when TBFs are used. One of the minor changes is that the log marginal likelihood values reported for the models correspond to the log Bayes factors versus the null model containing only an intercept term. Moreover, the parameter sampling works without MCMC, instead the coefficient samples are exactly simulated from an approximate mixture of Gaussian distributions (see Section 3.3.3 in Paper IV for details). The mixture is generated by the posterior distribution of g. If the conjugate prior was used, then the g samples

are exactly simulated from the updated distribution. Otherwise, the g samples are generated from a numerical approximation of the posterior density.

1.2 Splines

In order to use splines instead of FPs for modelling the covariate effects, we will use the methodology presented in Paper II and implemented in the R-package `hypergsplines`.

Model specification The first step is to define a `modelData` object, where we input the response vector `y`, the matrix with the covariates `X`, the spline type (here we use `cubic` O'Sullivan splines with 4 inner knots) and the exponential family (here a canonical `binomial` model, i.e. we want logistic regression):

```
library(hypergsplines)
modelData.pima <- with(pima, glmModelData(y = hasDiabetes, X = cbind(npreg,
    glu, bp, skin, bmi, ped, age), splineType = "cubic", nKnots = 4L,
    family = binomial))
```

By default, the hyper-g/n prior is specified as the hyperprior for g, via the S4 class constructor function `HypergnPrior`. It is passed via the `gPrior` option of `glmModelData`. The alternatives are the hyper-g prior (`HypergPrior`), the inverse-gamma prior (`InvGammaGPrior`) and any hyperprior with `CustomGPrior`.

Moreover, by default all covariates in the design matrix `X` are assumed to be continuous and to be modelled with splines. This is specified with the argument `continuous`. If you do have dummy variables or other covariates which you do not want to model with splines, then you set the corresponding element of the logical vector passed to `continuous` to `FALSE`.

Model prior and stochastic search Next, we will do a stochastic search on the (very large) model space to find "good" models. Here we have to decide on the model prior, and in this example we use the `dependent` type which corrects for the implicit multiplicity of testing. The alternatives are documented in the help page `?getLogModelPrior`.

```
chainlength.pima <- 100
computation.pima <- getComputation(useOpenMP = FALSE, higherOrderCorrection = FALSE,
    verbose = FALSE)
```

We use a `chainlength` of 100, which is very small but enough for illustration purposes (usually one should use at least 100 000), and save all models (in general `nModels` is the number of models which are saved from all visited models). Finally, we decide that we do not want to use OpenMP acceleration and no higher order correction for the Laplace approximations, and omit the progress bars. In order to be able to reproduce the analysis, it is advisable to set a seed for the random number generator before starting the stochastic search:

```
set.seed(93)
time.pima <- system.time(models.pima <- stochSearch(modelData = modelData.pima,
    modelPrior = "dependent", chainlength = chainlength.pima,
    nModels = chainlength.pima, computation = computation.pima))

##
##
## Number of non-identifiable model proposals:     0
## Number of total cached models:                  77
## Number of returned models:                      77

time.pima

##    user  system elapsed
##  17.944   0.096  18.109

models.pima[["numVisited"]]

## [1] 77
```

We see that the search took 18 seconds, and 77 models were found. The `models` list element of `models.pima` gives the table of the found models, with their degrees of freedom for every covariate, the log marginal likelihood, the log prior probability, the posterior probability and the number of times that the sampler encountered that model:

```
head(models.pima$models)

##   npreg glu bp skin bmi ped age logMargLik logPrior    post hits
## 1     3   1  0    0   4   2   4     -242.2   -14.08 0.28528    2
## 2     3   1  0    0   4   2   3     -242.5   -14.08 0.22813    2
## 3     3   1  0    0   3   2   3     -242.5   -14.08 0.21329    2
## 4     1   1  0    0   2   2   2     -244.3   -14.08 0.03690    5
## 5     3   1  1    0   4   2   4     -243.7   -14.78 0.03387    3
## 6     0   1  0    3   4   2   3     -244.7   -14.08 0.02576    0

map.pima <- models.pima$models[1, 1:7]
```

Note that we just saved the degrees of freedom vector of the estimated MAP model in `map.pima`.

Inclusion probabilities The estimated marginal inclusion probabilities (probabilities for exclusion, linear inclusion and nonlinear inclusion) for all covariates are also saved:

```
round(models.pima$inclusionProbs, 3)

##              npreg   glu    bp  skin   bmi   ped age
## not included 0.031 0.000 0.908 0.861 0.000 0.000   0
## linear       0.069 0.955 0.092 0.089 0.037 0.015   0
## non-linear   0.900 0.045 0.000 0.050 0.963 0.985   1
```

These probabilities are based on the renormalised model probabilities of all models encountered during the stochastic model search. In this toy example, we saved all models, so there is no difference to

```
round(getInclusionProbs(models.pima$models, modelData = modelData.pima),
    3)

##              npreg   glu    bp  skin   bmi   ped age
## not included 0.031 0.000 0.908 0.861 0.000 0.000   0
## linear       0.069 0.955 0.092 0.089 0.037 0.015   0
## non-linear   0.900 0.045 0.000 0.050 0.963 0.985   1
```

which gives the probabilities based on all saved models.

Sampling model parameters If we now want to look at the estimated covariate effects in the estimated MAP model which has configuration (3,1,0,0,4,2,4), then we first need to generate parameter samples from that model:

```
mcmc.pima <- getMcmc(samples = 500L, burnin = 100L, step = 1L,
    nIwlsIterations = 2L)
set.seed(634)
map.samples.pima <- glmGetSamples(config = map.pima, modelData = modelData.pima,
    mcmc = mcmc.pima, computation = computation.pima)
```

With the function `getMcmc`, we have defined a list of MCMC settings, comprising the number of samples we would like to have in the end, the length of the burn-in, the thinning step (here no thinning) and the number of IWLS iterations used (with 2 steps you get a higher acceptance rate than with 1 step, here the acceptance rate was 0.51). The result `map.samples.pima` has the following structure:

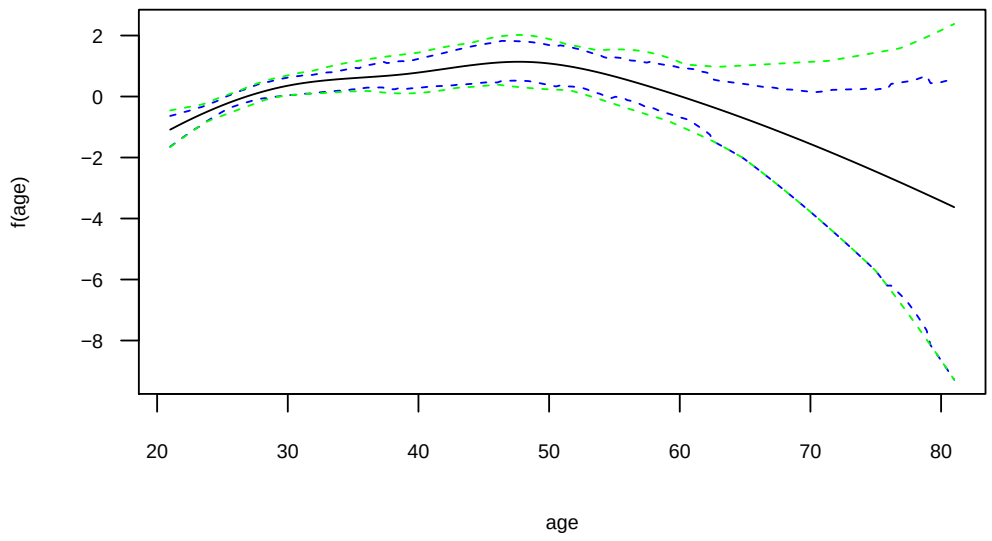
```
str(map.samples.pima)

## List of 3
## $ samples :List of 5
## ..$ t : num [1:500] 0.992 0.992 0.992 0.992 0.992 ...
## ..$ intercept : num [1:500(1d)] -0.986 -0.986 -0.986 -0.986 -0.986
##    ...
## ..$ linearCoefs:List of 5
## .. ..$ npreg: num [1:500(1d)] -0.397 -0.397 -0.397 -0.397 -0.397 ...
## .. ..$ glu : num [1:500(1d)] 5.98 5.98 5.98 5.98 5.98 ...
## .. ..$ bmi : num [1:500(1d)] 4.54 4.54 4.54 4.54 4.54 ...
## .. ..$ ped : num [1:500(1d)] 4.02 4.02 4.02 4.02 4.02 ...
## .. ..$ age : num [1:500(1d)] 3.3 3.3 3.3 3.3 3.3 ...
## ..$ splineCoefs:List of 4
## .. ..$ npreg: num [1:6, 1:500] 12.09 -5.8 13.93 -5.88 1.91 ...
## .. ..$ bmi : num [1:6, 1:500] 5.03 -22.3 56.31 -17.7 -30.82 ...
## .. ..$ ped : num [1:6, 1:500] -6.45 7.46 1.59 -10.33 -7.07 ...
## .. ..$ age : num [1:6, 1:500] 48.494 26.44 3.673 -0.356 -12.227 ...
## ..$ z : num [1:500] 4.87 4.87 4.87 4.87 4.87 ...
## $ mcmc :List of 2
## ..$ nAccepted : num 306
## ..$ acceptanceRatio: num 0.51
## $ logMargLik:List of 4
```

```
## ..$ ilaEstimate : num -242
## ..$ mcmcEstimate : num -242
## ..$ mcmcSe : num 0.0867
## ..$ margApproxZdens:List of 2
## .. ..$ args: num [1:100] -52.0976 -29.1796 -0.8346 -0.0443 0.4362 ...
## .. ..$ dens: num [1:100] 0.00 1.41e-51 7.72e-30 3.36e-23 1.09e-18 ...
```

It is a list with the `samples`, two diagnostics for the `mcmc`, and estimates for the log marginal likelihood (`logMargLik`). The latter one contains the original integrated Laplace approximation (ILA) estimate, the MCMC estimate of the log marginal likelihood with its standard error, and the coordinates of the posterior density of $z = \log(g)$.

Curve estimates Now we can use the samples to plot the estimated effects of the MAP model covariates, with the `plotCurveEstimate` function. For example:

```
plotCurveEstimate(covName = "age", samples = map.samples.pima$samples,
    modelData = modelData.pima)
```



Instead of the mean curve (plotted in black), also the median curve could be plotted here, by specifying the option `estimate="median"` in `plotCurveEstimate`. Analo-

gously to the FP curves, pointwise 95% credible intervals are plotted in blue, and a simultaneous 95% credible band is plotted in green.

Post-processing If we want to have estimates of the degrees of freedom on a continuous scale instead of the fixed grid $(0, 1, 2, 3, 4)$, we can optimise the marginal likelihood with respect to the degrees of freedom of the MAP covariates:

```
optim.map.pima <- postOptimize(modelData = modelData.pima, modelConfig = map.pima,
    computation = computation.pima)
optim.map.pima

##   npreg glu bp skin   bmi   ped   age
## 1 2.277   1  0    0 3.556 2.104 3.654
```

For that model, we could again produce samples and plot curve estimates.

Prediction samples If we would like to get prediction samples for new covariate values, this is also very easy via the `getFitSamples` function. Here we get posterior predictive samples because we input a covariate matrix which is part of the original covariate matrix used to fit the MAP model. Because the `getFitSamples` function produces samples on the linear predictor scale, we have to apply the appropriate response function (here the logistic distribution function `plogis`) to get samples on the observation scale.

```
fit.samples.pima <- getFitSamples(X = modelData.pima$origX[1:10,
    ], samples = map.samples.pima$samples, modelData = modelData.pima)
obs.samples.pima <- plogis(fit.samples.pima)
```

The posterior predictive means are thus:

```
rowMeans(obs.samples.pima)

##  [1] 0.04662 0.64899 0.08634 0.69537 0.03228 0.28551 0.06224 0.59718
##  [9] 0.27026 0.67184
```

and could be compared to the actual observations

```
modelData.pima$Y[1:10]

##  [1] 0 1 0 0 0 1 0 0 0 1
```

Model averaging Model averaging works in principle similar to sampling from a single model, but multiple model configurations are supplied and their respective log posterior probabilities. For example, if we wanted to average the top ten models found, we would do the following:

```
average.samples.pima <- with(models.pima, getBmaSamples(config = models[1:10,
    ], logPostProbs = log(models$post[1:10]), nSamples = 500L,
    modelData = modelData.pima, mcmc = mcmc.pima, computation = computation.pima))
```

Then internally, first the models are sampled, and for each sampled model so many samples are drawn as determined by the model frequency in the model average sample. On this sample object, the above presented functions can again be applied (e.g. `plotCurveEstimate`).

2 Cox Regression

Now we would like to illustrate the analysis of survival data with Cox regression (see Section 1 in the Introduction). We consider survival data provided by Therneau and Grambsch (2000) on primary biliary cirrhosis (PBC) patients, from which we use the $n = 276$ complete observations. The data is contained in the R-package `survival` (Therneau and Grambsch, 2000):

```
library(survival)
## restrict to full observations
pbcFull <- na.omit(pbc)
pbc.nObs <- nrow(pbcFull)
## introduce censoring indicator: only if death (status == 2)
## is observed, the total survival time is observed.
## Transplantation is also a censoring!
pbcFull$observed <- pbcFull$status == 2
pbc.nEvents <- sum(pbcFull$observed)
```

```r
## encode factors
library(Hmisc)
pbcFull <- upData(pbcFull, levels = list(trt = list(penicillamin = 1,
    placebo = 2), sex = list(male = "m", female = "f"), ascites = list(no = 0,
    yes = 1), hepato = list(no = 0, yes = 1), spiders = list(no = 0,
    yes = 1), edema = list(no = 0, notTreatedOrSuccessful = 0.5,
    therapyResistant = 1), stage = list(`1` = 1, `2` = 2, `3` = 3,
    `4` = 4)))

## Input object size:  40016 bytes;  21 variables
## New object size:  40712 bytes;  21 variables

head(pbcFull)

##   id time status          trt   age    sex ascites hepato spiders
## 1  1  400      2 penicillamin 58.77 female     yes    yes     yes
## 2  2 4500      0 penicillamin 56.45 female      no    yes     yes
## 3  3 1012      2 penicillamin 70.07   male      no     no      no
## 4  4 1925      2 penicillamin 54.74 female      no    yes     yes
## 5  5 1504      1      placebo 38.11 female      no    yes     yes
## 7  7 1832      0      placebo 55.53 female      no    yes      no
##                    edema bili chol albumin copper alk.phos    ast
## 1       therapyResistant 14.5  261    2.60    156     1718 137.95
## 2                     no  1.1  302    4.14     54     7395 113.52
## 3 notTreatedOrSuccessful  1.4  176    3.48    210      516  96.10
## 4 notTreatedOrSuccessful  1.8  244    2.54     64     6122  60.63
## 5                     no  3.4  279    3.53    143      671 113.15
## 7                     no  1.0  322    4.09     52      824  60.45
##   trig platelet protime stage observed
## 1  172      190    12.2     4     TRUE
## 2   88      221    10.6     3    FALSE
## 3   55      151    12.0     4     TRUE
## 4   92      183    10.3     4     TRUE
## 5   72      136    10.9     3    FALSE
## 7  213      204     9.7     3    FALSE

## save the covariate names
pbcCovNames <- setdiff(names(pbcFull), c("id", "time", "status",
    "observed"))
```

2.1 Fractional Polynomials

As hyperprior on g, we choose the hyper-g/n_{obs} prior, which was proposed in Paper IV:

```
prior.hypergn <- CustomGPrior(logDens = function(g) {
    return(-log(pbc.nEvents) - 2 * log1p(g/pbc.nEvents))
})
```

For the model formula, we just need to know which covariates are factors and which covariates are continuous variables:

```
formula.pbc <- time ~ uc(trt) + bfp(age) + uc(sex) + uc(ascites) +
    uc(hepato) + uc(spiders) + uc(edema) + bfp(bili) + bfp(chol) +
    bfp(albumin) + bfp(copper) + bfp(alk.phos) + bfp(ast) + bfp(trig) +
    bfp(platelet) + bfp(protime) + uc(stage)
```

The response on the left-hand side of the formula is the survival time.

When analysing survival data with the R-package `glmBfp`, it is recommended to use TBFs, because they do not require data augmentation (as it is needed for the splines in Section 2.2) and are thus much faster. In fact, only TBFs are supported with survival data for `glmBfp`, so the data augmentation as discussed in Paper III would have to be done manually if it is needed.

We start the model search as in the following:

```
set.seed(911)
models.pbc <- glmBayesMfp(formula.pbc, censInd = pbcFull$observed,
    data = pbcFull, tbf = TRUE, priorSpecs = list(gPrior = prior.hypergn,
        modelPrior = "dependent"), nModels = 1000, chainlength = 100,
    method = "sampling", verbose = FALSE)

## Warning:  Input data were reordered so that the survival times are sorted
```

It is important to set `tbf=TRUE` for use of the TBFs, and to pass the logical censoring indicator vector to the `censInd` argument of `glmBayesMfp`. If the survival time has been fully observed, the corresponding element of `censInd` is `TRUE`, if is censored it is `FALSE`. When this argument is used, the program knows that Cox regression should be performed. Note the warning that the input data were reordered. This can be

observed in the attribute `data` of the result, where the rownames of the design matrix tell the new order:

```
head(attr(models.pbc, "data")$x)

##     (Intercept)    age albumin alk.phos   ast bili  chol copper
## 281           1 0.6588   0.210    0.705 3.380 1.79 0.175   2.20
## 10            1 0.7056   0.274    0.918 1.472 1.26 0.200   1.40
## 76            1 0.5194   0.308    2.132 1.550 1.22 0.394   1.11
## 27            1 0.5444   0.331    3.697 1.019 2.16 0.175   2.21
## 103           1 0.4896   0.367    1.273 1.194 0.25 0.188   0.57
## 18            1 0.5393   0.280    0.961 2.806 1.14 0.178   5.88
##     platelet protime trig ascitesyes edemanotTreatedOrSuccessful
## 281     0.62    1.29 2.29          1                           0
## 10      3.02    1.15 1.43          1                           0
## 76      1.65    1.16 2.43          0                           1
## 27      0.80    1.20 1.68          1                           1
## 103     1.10    1.11 1.02          1                           0
## 18      2.83    1.24 2.00          0                           0
##     edematherapyResistant hepatoyes sexfemale spidersyes stage2
## 281                     1         0         1          0      0
## 10                      1         0         1          1      0
## 76                      0         1         1          1      0
## 27                      0         1         1          1      0
## 103                     1         1         1          1      0
## 18                      1         1         1          1      0
##     stage3 stage4 trtplacebo
## 281      0      1          0
## 10       0      1          1
## 76       0      1          0
## 27       0      1          1
## 103      0      1          1
## 18       0      1          0
```

This must be taken into account when interpreting *e.g.* the fitted linear predictor values. It may thus be advisable to instead manually order the data before running `glmBayesMfp`.

All other functionality described for logistic regression in Section 1 can be applied in the same way here for Cox regression. For example, we can obtain the inclusion probabilities via

```
round(attr(models.pbc, "inclusionProbs"), 3)

##      age  albumin alk.phos      ast     bili     chol   copper
##    0.933    0.980    0.009    0.000    1.000    0.000    0.010
## platelet  protime     trig  ascites    edema   hepato      sex
##    0.001    0.786    0.001    0.145    0.000    0.000    0.000
##  spiders    stage      trt
##    0.000    0.000    0.000
```

The only minor difference is that there is no intercept term in the Cox model (or, to be more precise, it is not estimated here), so the corresponding samples are missing in the samples objects.

2.2 Splines

Using the Poisson approximation of the Cox regression model described in Paper III, we can also apply the generalised additive model selection methodology from Paper II to survival data, which is also implemented in the R-package `hypergsplines`.

Model specification As in Section 1.2, we first have to define a design matrix, because there is currently no formula interface for model specification:

```
## form design matrix
X <- model.matrix(time ~ trt + age + sex + ascites + hepato +
    spiders + edema + bili + chol + albumin + copper + alk.phos +
    ast + trig + platelet + protime + stage, data = pbcFull)[,
    -1]
head(X)

##   trtplacebo   age sexfemale ascitesyes hepatoyes spidersyes
## 1          0 58.77         1          1         1          1
## 2          0 56.45         1          0         1          1
## 3          0 70.07         0          0         0          0
## 4          0 54.74         1          0         1          1
## 5          1 38.11         1          0         1          1
## 7          1 55.53         1          0         1          0
##   edemanotTreatedOrSuccessful edematherapyResistant bili chol albumin
## 1                           0                     1 14.5  261    2.60
```

```
## 2                          0                    0  1.1  302    4.14
## 3                          1                    0  1.4  176    3.48
## 4                          1                    0  1.8  244    2.54
## 5                          0                    0  3.4  279    3.53
## 7                          0                    0  1.0  322    4.09
##   copper alk.phos    ast trig platelet protime stage2 stage3 stage4
## 1    156     1718 137.95  172      190    12.2      0      0      1
## 2     54     7395 113.52   88      221    10.6      0      1      0
## 3    210      516  96.10   55      151    12.0      0      0      1
## 4     64     6122  60.63   92      183    10.3      0      0      1
## 5    143      671 113.15   72      136    10.9      0      1      0
## 7     52      824  60.45  213      204     9.7      0      1      0
```

For this data set, not all covariates are continuous, therefore we have to specify which of them are:

```
pbc.cont <- c(FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE,
    FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,
    FALSE, FALSE, FALSE)
## for checking the correct matching between covariates and
## options:
cbind(colnames(X), pbc.cont)

##                                     pbc.cont
##  [1,] "trtplacebo"                  "FALSE"
##  [2,] "age"                         "TRUE"
##  [3,] "sexfemale"                   "FALSE"
##  [4,] "ascitesyes"                  "FALSE"
##  [5,] "hepatoyes"                   "FALSE"
##  [6,] "spidersyes"                  "FALSE"
##  [7,] "edemanotTreatedOrSuccessful" "FALSE"
##  [8,] "edematherapyResistant"       "FALSE"
##  [9,] "bili"                        "TRUE"
## [10,] "chol"                        "TRUE"
## [11,] "albumin"                     "TRUE"
## [12,] "copper"                      "TRUE"
## [13,] "alk.phos"                    "TRUE"
## [14,] "ast"                         "TRUE"
## [15,] "trig"                        "TRUE"
## [16,] "platelet"                    "TRUE"
```

```
## [17,] "protime"                    "TRUE"
## [18,] "stage2"                     "FALSE"
## [19,] "stage3"                     "FALSE"
## [20,] "stage4"                     "FALSE"
```

Then finally, we can define the model space, using the `survModelData` function:

```
## now do data augmentation for the Poisson approximation
library(hypergsplines)
pbc.md <- survModelData(times = pbcFull$time, X = X, observed = pbcFull$observed,
    continuous = pbc.cont, nKnots = 4L, splineType = "cubic",
    gPrior = HypergnPrior(a = 4, n = pbc.nEvents))

## Warning:  Input data were reordered so that the survival times are sorted
```

The survival times are passed in the `times` argument, and the censoring indicator in the `observed` argument. Note that we use again cubic splines with 4 knots, and the hyper-g/n_{obs} prior for g. The output object `pbc.md` (not of a formal class) can now be used exactly in the same way as we used the `modelData.pima` object in Section 1.2, because it is a Poisson regression object.

For example, we can do a stochastic search:

```
set.seed(27)
time.pbc <- system.time(models.pbc <- stochSearch(modelData = pbc.md,
    modelPrior = "dependent", chainlength = 10, nModels = 10,
    computation = computation.pima))

##
##
## Number of non-identifiable model proposals:     0
## Number of total cached models:                  9
## Number of returned models:                      9

time.pbc

##    user  system elapsed
##  65.720   4.092  70.034

models.pbc[["numVisited"]]

## [1] 9
```

We note that the computations take much longer here, with 70 seconds for 9 models in this example. This is due to the large sample size in the augmented data set:

```
pbc.md$nObs

## [1] 37155
```

References

F. Colavecchia and G. Gasaneo. f1: a code to compute Appell's F1 hypergeometric function. *Computer Physics Communications*, 157(1):32–38, 2004.

R. C. Forrey. Computing the hypergeometric function. *Journal of Computational Physics*, 137(1):79–100, 1997.

A. Frank and A. Asuncion. UCI Machine Learning Repository, 2010. URL http://archive.ics.uci.edu/ml.

G. García-Donato and M. A. Martinez-Beneito. On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481): 410–423, 2008.

N. Michel and M. Stoitsov. Fast computation of the Gauss hypergeometric function with all its parameters complex with application to the Pöschl–Teller–Ginocchio potential wave functions. *Computer Physics Communications*, 178(7):535–551, 2008.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11, 2006.

B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

D. Sabanés Bové and L. Held. Hyper-*g* priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410, 2011a.

D. Sabanés Bové and L. Held. Bayesian fractional polynomials. *Statistics and Computing*, 21(3):309–324, 2011b.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.